Power of Alternative Fit Indices for Multiple Group Longitudinal Tests of Measurement Invariance

By

Stephen D. Short

Submitted to the Department of Psychology and the
Graduate Faculty of the University of Kansas
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Pascal R. Deboeck, Chairperson

Todd D. Little, Co-Chairperson

Committee members

Carol M. Woods

Wei Wu

William P. Skorupski

Date defended: April 14, 2014

The Dissertation Committee for Stephen D. Short certifies
that this is the approved version of the following dissertation :

Power of Alternative Fit Indices for Multiple Group Longitudinal Tests of Measurement Invariance

Pascal R. Deboeck, Chairperson

Date approved: April 17, 2014

Abstract

Measurement invariance testing with confirmatory factor analysis has a long history in social science research, and more recently has increased use and popularity. The current paper begins by reviewing the steps for measurement invariance testing via multiple group confirmatory factor analysis, and synthesizing previous research recommendations for model testing, including the chi-square difference test, and examining change in model fit indices. Previous research on measurement invariance testing has examined change in alternative fit indices such as the *CFI*, *TLI*, *RMSEA*, and *SRMR*, but these studies had not examined power to detect invariance when more than two groups exist and multiple time points are present. The present study implemented a Monte Carlo simulation to examine the power of change in alternative fit indices to detect two types of measurement invariance, weak and strong, across a variety of manipulated study conditions including sample size, sample size ratio, lack of invariance, location of noninvariance, magnitude of noninvariance, and type of mixed study design.

Acknowledgements

First, I want to thank each of my committee members: Dr. Pascal Deboeck, Dr. Todd Little, Dr. Wei Wu, Dr. Carol Woods, and Dr. William Skorupski. I greatly appreciate your time providing me with feedback. I also would like to thank my colleagues and friends for all of their support.

Lastly, I especially would like to thank my family. I can never truly express how thankful I am for the love and support from my fiancé Chelsea Reid, my twin sister Ashley Borgie, and my parents, David and Teresa Short, who always offered encouragement and listened in times of stress. Over the past five years I have come to love KU and Lawrence, KS, but I am so very excited to soon no longer be halfway across the country from the most important individuals in my life. I dedicate this dissertation to each of you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Psychological researchers are frequently interested in examining differences between groups, such as gender, nationality, ethnicity, or culture. Underlying these examinations is typically the assumption that the measure being used for comparisons functions the same across groups. For example, a social psychologist may be interested in studying differences in self-enhancement, which is defined as the tendency to have positive views of one's self (see Heine & Hamamura, 2007). A scale designed to measure an individual's self-enhancement may function differently for an individual from a western, individualistic culture, than for an individual from an eastern, collectivist culture. If a researcher were to make comparisons in self-enhancement between these cultures, differences may exist due to true cross-cultural differences, or simply differences in the measurement scale's properties. Thus, social science researchers interested in making group comparisons may first want to examine if the properties of their measure are invariant across groups. Measurement invariance has been previously reviewed by a number of authors (e.g., Alwin & Jackson, 1981; French & Finch, 2006; Little, 1997, 2013; Marsh, 1994; Meredith, 1964, 1993; Meredith & Horn, 2001; Millsap, 2011; Reise et al., 1993; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000). The current study serves to synthesize the literature, present recent research findings, and contribute new findings from a simulation study.

1.1 A Brief History of Measurement Invariance

Simply stated, latent variable models postulate that a participant's observed (i.e., manifest) responses to a set of measured items are caused by one or more unobservable latent variables. For example, an individual's response to several items on a mathematics exam may be caused by a latent quantitative ability construct. Measurement invariance is defined as individuals from different groups or time points having equal conditional probabilities of having a certain observed score, given that they share the same score on the underlying latent construct (Meredith, 1993). If measurement invariance is not established, then measurement bias is said to be present and examinations of group or time differences in the latent variable may be compromised (Millsap & Olivera-Aguilar, 2012). Thus, the examination of measurement invariance can help determine if examining the effect of group or time is warranted, or if change is occurring in the properties of a latent construct (Widaman et al., 2010).

The study of measurement invariance is deeply rooted in the history of factor analysis. Measurement invariance in factor analysis most frequently examines differences across groups, thus, the foundation of measurement invariance research begins in the study of selection theory, which states how observations from a population can be assigned into different groups (Millsap et al., 2007). Aitken (1935) discussed the impact of selection processes on covariances structures for measured items between groups, which Thurstone (1947) expanded upon by demonstrating that even after selection occurs and various groups, or subpopulations exist, simple structure (e.g., each measured variable has a large loading on one factor and near zero loadings on all other factors) can be maintained across groups in exploratory factor analyses (EFA) results.

Meredith (1964) further supported Thurstone's (1947) claim by demonstrating the simple structure solution from an EFA conducted for each subpopulation could be set to be invariant across the subpopulations. Following Meredith (1964), research continued to focus on various rotational methods in EFA that can lead to invariant factor loadings (Millsap et al. 2007). However, Jöreskog (1971) presented the ability to simultaneously estimate confirmatory factor analysis (CFA) models across groups, allowing researchers the ability to begin comparing a CFA model where parameter

estimates could be both freely estimated or constrained (i.e., invariant) across groups. Sörbom (1974) extended this work to the means structures model within CFA, thus allowing researchers to begin examining not only equivalence across the covariances structure (e.g., factor loadings, factor variances, and residuals), but also the item intercepts and factor means.

Although the foundation for examining measurement invariance with factor analysis was established throughout the middle of the 20th century, examinations of the analytic technique increased during the last 30 years. For example, the use of multiple group CFA for testing measurement invariance was discussed for mulitrait-multimethod data (Cole & Maxwell, 1985), experimental designs (Bagozzi, 1977; Bagozzi et al., 1991; Hancock, 1997), and factorial designs (Marsh, 1994). Many applications of measurement invariance testing focused on between groups differences on demographic variables, such as gender (Crawford & Henry, 2004), ethnicity (Glanville & Wildhagen, 2007), age (Horn & McArdle, 1992), or culture (Little, 1997), but invariance can also be examined across time (Coertjens et al., 2012; McArdle, 2009; Meredith & Horn, 2001; Pentz & Chou, 1994; Widaman et al., 2010) or both group and time (Little, 1997; Raykov, 2005). A researcher may be interested in establishing measurement invariance in order to support the validity of the measure across populations, or as a necessary first step before investigating other potential differences between groups or time, such as factor means (Ployhart & Oswald, 2004), factor variances, or factor covariances.

1.2 Common Types of Measurement Invariance with CFA

Researchers investigating measurement invariance with CFA can test many different hypotheses which can be summarized as tests examining factor structure, factor loadings, item intercepts, item unique variances (i.e., residual variance), factor variances/covariances, and factor means (see Cheung & Rensvold, 1999; Little, 1997; Vandenberg & Lance, 2000). Traditionally, an omnibus test of equivalence for the estimated covariance matrix between groups has been suggested as a first step in examining measurement invariance (Alwin & Jackson, 1981; Byrne et al., 1989; Jöreskog,

1971; Steenkamp & Baumgartner, 1998). If the covariance matrices are found to be equivalent across groups, (i.e., the relationships between measured variables are the same across groups), then measurement invariance is said to be present. Unfortunately, the omnibus test of homogenous variance/covariance matrices for each group can be very difficult to pass, as any differences in factor loadings, factor variances, and residual variances will likely lead to a significant difference between the groups (Vandenberg & Lance, 2000). If a difference between the matrices is detected, then a series of follow-up of tests to determine where noninvariance (i.e., measurement bias) may exist is necessary. Meredith (1993) advanced the discussion of measurement invariance testing in latent variables via multiple group CFA, including the processes of examining loading invariance (e.g., weak factorial invariance), intercept invariance (e.g., strong factorial invariance), and residual invariance (e.g., strict factorial invariance). Researchers examining measurement invariance across several time points would follow the same conditions (see Widaman et al. 2010), and could examine measurement invariance across both group and time simultaneously (see Little, 2013). Following Meredith, 's (1993) terminology, configural, weak, strong, and strict factorial invariance are discussed in more detail below. Readers should note that there are some discrepancies with terminology used in the past research on examining measurement invariance with CFA. For example, Vanderberg and Lance (2000) refer to configural invariance as "weak" invariance and weak invariance as "strong" invariance.

1.2.1 Configural invariance

In the configural invariance model the same pattern of fixed and freely estimated parameters is specified across each group for both the mean and covariance structure. In other words, each group has the same measurement model specified. Equation 1.1 and 1.2 display the covariance and means structure configural invariance model.

$$\Sigma_{gt} = \Lambda_{gt}\Psi_{gt}\Lambda'_{gt} + \Theta_{gt} \tag{1.1}$$

$$\mu_{ygt} = T_{gt} + \Lambda_{gt} A_{gt} \tag{1.2}$$

Where $_g$ = group, $_t$ = time, Σ_{gt} = estimated covariance matrix for group $_g$ at time $_t$, Λ_{gt} = estimated factor loading matrix for group $_g$ at time $_t$, Θ_{dgt} = estimated residual covariance matrix for group $_g$ at time $_t$, μ_{ygt} = a vector of estimated means for item $_y$ in group $_g$at time $_t$, T_{gt} = a vector of estimated intercepts for item $_y$ in group $_g$ in time $_t$, and A_{gt} = estimated latent factor means for group $_g$ at time $_t$. Thus, in the configural invariance model the estimated parameters do not need to be equal across group or time, but simply share the same factor structure pattern

1.2.2 Weak invariance

Weak invariance has also been referred to as metric invariance (Horn & McArdle, 1992; Millsap & Olivera-Aguilar, 2012) or factor loading invariance and is a test of factor loading equivalence across groups and/or time. Equations 1.3 and 1.4 demonstrate weak invariance.

$$\Sigma_{gt} = \Lambda \Psi_{gt} \Lambda + \Theta_{gt} \tag{1.3}$$

$$\mu_{ygt} = T_{gt} + \Lambda A_{gt} \tag{1.4}$$

In the weak invariant model the estimated factor loading matrix (Λ) for each group is constrained to be equal, so that only one set of factor loadings are estimated across groups and/or time. The newly specified weak invariant model's fit statistics are then compared back to the configural model to test the null hypothesis that the factor loadings are equal across groups. The statistical tests used to evaluate weak invariance are discussed the next section. In general, the test of weak invariance can be considered a hypothesis test where the null hypotheses states that the factor loadings are equivalent. Therefore, if the researcher fails to reject the null hypothesis, then weak invariance is established. Confirmed weak invariance demonstrates that the same mea-

sured items across each group and/or time point have equal amounts of variability explained by the latent construct. A researcher must establish weak invariance prior to investigating potential differences in factor variances and covariance matrix, Ψ (Millsap & Olivera-Aguilar, 2012; Schmitt & Kuljanin, 2008)

1.2.3 Strong invariance

Strong invariance may also be referred to as scalar (Millsap & Olivera-Aguilar, 2012) or intercept invariance and involves both constraining factor loadings and item intercepts to equality across groups and/or time. Equations 1.5 and 1.6 depict the strong invariance model.

$$\Sigma_{gt} = \Lambda\Psi_{gt}\Lambda + \Theta_{gt} \tag{1.5}$$

$$\mu_{ygt} = T_y + \Lambda A_{gt} \tag{1.6}$$

In other words, across each group and time point, the same items are said have the same scores for each person who has a score of zero on the latent construct. Strong invariance is necessary before researcher is able to make comparisons of latent means. Similar to evaluation of the weak invariance model, where weak invariance is evaluated only after configural invariance is established, the strong invariant model is examined only after weak invariance is established. The null hypothesis that the item intercepts are equal across groups is tested by comparing the restricted (i.e., constrained) strong invariance model to the full (i.e., freely estimated) weak invariant model. Interestingly, the investigation of strong invariance may often be ignored in applied research. For example, Vanderberg and Lance (2000) note that contrary to weak invariance, strong invariance does not appear to frequently be examined in psychological research, particularly in the area of I/O psychology. A more recent study by Schmitt and Kuljanin (2008) reports this examination is increasing in I/O psychology. Strong invariance is required before researchers can proceed to testing latent means (Hancock, 2001)

1.2.4 Strict invariance

Strict invariance may also be referred to as residual invariance and involves not only constraining the factor loadings and item intercepts to equality, but also the residual (i.e., unique) variances. Equations1.7 and 1.8 detail the strict invariance model.

$$\Sigma_{gt} = \Lambda \Psi_{gt} \Lambda + \Theta \tag{1.7}$$

$$\mu_{ygt} = T_y + \Lambda A_{gt} \tag{1.8}$$

Some debate exists around the examination of strict invariance. DeShon (1998, 2004) reminds researchers that the unique variances estimated in CFA contain both random, uncorrelated error variance, and also specific variance due to other factors not modeled. DeShon (2004) proposes that a lack of strict factorial invariance demonstrates the same latent construct may be influenced differently by one or more extraneous, unmodeled constructs, making comparisons between groups inaccurate. Conversely, in practice many researchers do not view strict invariance as a necessary requirement for establishing measurement invariance (Little, 2013; Millsap & Olivera-Aguilar, 2012). Equating unique item variances across groups has been suggested as an unreasonable constraint that frequently will not hold when analyzing real data (Dimitrov, 2010; Little, 2013; Schmitt & Kuljanin, 2008; Widaman & Reise, 1997).

1.2.5 Partial invariance

When a null hypothesis for one of the tests of measurement invariance is rejected, a researcher may be interested in determining which item or items are noninvariant across groups or time. Partial invariance exists when some, but not all, of the items in a given measure are invariant across groups. Partial invariance can exist at the weak, strong, or strict invariance level. Byrne et al. (1989) first introduced the concept of establishing partial invariance, and provided an example demonstrating their recommended procedure. When a researcher fails to establish weak or strong

invariance, Byrne et al. (1989) suggest that modification indices for the model could be viewed for each equality constraint. The largest modification index for relaxing an equality constraint could be freed, and the researcher could then determine if freeing the parameter was enough to then satisfy the conditions for their desired level of invariance. This process could continue until the desired level of invariance is achieved. Alternatively, a researcher could constrain a single item parameter across groups to equality and examine if the constraint is tenable. If so, the researcher could proceed with each of the remaining items, until the noninvariant items are located. Certainly, a researcher implementing statistical tests for this examination should consider a correction for the inflated Type I error rate (Millsap, 2011).

In addition, researchers interested in establishing partial invariance should be aware of a few more concerns. First, as soon as the researcher begins examining modification indices for suggestions on which items may be noninvariant across groups, the examination begins to be driven more by data than theory. Although a modeling approach driven by both data and theory may be necessary, a researcher is still encouraged to have substantive reasons for why an item may be noninvariant (Byrne et al., 1989). A strictly data-driven approach may not be replicable in future studies. For example, different studies examining invariance across populations may develop several "short forms" or subsets of the original measure, resulting in several measures that may not be directly comparable (Cheung & Rensvold, 1998). Second, determining partial invariance can potentially be a laborious search for noninvariant items. Simulations studies reveal that when the overall percentage of noninvariant items small, post-hoc searches using modification indices can be effective in detecting these items, but when noninvariant items are the majority in a factor model these post-hoc examinations can fail to detect which items are noninvariant (Yoon & Millsap, 2007). Millsap & Kwok (2004) examined the effects of using the all of the scale items, ignoring the noninvariant (i.e., the existence of partial measurement invariance) for a single factor construct for the purposes of selecting different populations, concluding that as both weak and strong partial invariance decreased (e.g., less equal factor loadings and intercepts across groups) sensitivity for selection dramatically decreased.

Latent mean invariance testing may be permitted as long as partial weak and strong measurement invariance is established (Hancock et al., 2000). However, an important question remains. If the researcher is able to specify partial invariance, how much invariance is required at the weak, or strong levels to proceed to testing latent means? Dimitrov (2010) notes that the answer to this question still appears to be subjective and suggests not exceeding 20%. Similarly, Sass (2011) states that partial measurement invariance may be suitable for proceeding to latent means testing, as long as the ratio of invariant to noninvariant items is large, but provides no empirical criteria for what constitutes a large ratio. Simply put, additional research on the merits of partial invariance is needed.

1.3 Evaluating Measurement Invariance with CFA

The conceptual framework for the various levels or hypotheses of measurement invariance testing has been defined, and the process for testing these hypotheses are now discussed below. First, the traditional chi-square test of nested model comparison is described and limitations are discussed. Then, several popular alternative model fit indices to $\chi^2$are defined, and previous research examining the use of these fit indices for testing measurement invariance hypothesis is described.

1.3.1 Chi-square nested model comparison

Earlier examinations of measurement invariance (e.g., Meredith, 1993) focused on a nested model comparison framework where the χ^2 obtained from the model with invariance constraints imposed was compared to the χ^2 from a model where the parameters were freely estimated. Specifically,

$$\Delta\chi^2 = \chi^2_{constrained} - \chi^2_{unconstrained} \tag{1.9}$$

$$\Delta df = df_{constrained} - df_{unconstrained} \tag{1.10}$$

where $\Delta\chi^2$ is itself χ^2 distributed and compared to a χ^2 distribution with Δdf degrees of freedom. If the $\Delta\chi^2$exceeds the critical value at the desired α level, then the model constraint is not supported and measurement bias exists. Simulations studies have revealed that testing measurement invariance via CFA and the χ^2 difference test is effective at detecting noninvariance, particularly as the number of items that are noninvariant increases, and the pattern of invariance is mixed (e.g., some parameters are higher in a focal group, and lower in the reference group, whereas other parameters are higher in reference group, but lower in the focal group; Meade & Lautenschlager, 2004).

When testing for weak invariance, the χ^2 difference test shows acceptable power (e.g., > .80) when communalities are high, the number of factors is low, and understandably, sample size per group is large (Meade & Bauer, 2007). However, as sample sizes increased, the χ^2 difference test may be overpowered and detect what researchers may consider trivial differences in measured items (Meade & Bauer, 2007). Indeed, the χ^2 difference test is often considered an overly sensitive test of measurement invariance that frequently suggests measurement bias exists when little is present (Brannick, 1995; Kelloway, 1995). Moreover, Brannick (1995) notes that as sample size (N) increases, the χ^2 difference test will eventually always be significant because χ^2 and N are dependent. Instead, recent research (e.g., Chen, 2007; Cheung & Rensvold, 1998; Meade et al., 2008) have focused on examining change in other model fit indices when testing for measurement invariance.

1.3.2 Alternative fit indices ($AFIs$)

In the current paper, alternative fit indices ($AFIs$) refers to other developed model fit indices that are alternatives to the estimated model χ^2. Throughout the history of CFA dozens of model fit indices have been developed and examined as useful tools for model evaluation. Researchers interested in a thorough review of alternative fit indices are encouraged to consult previous discussions (e.g., Brown, 2006; Hu & Bentler, 1999; Kline, 2011; Tanka, 1993). Alternative fit indices are used to examine measurement invariance following logic similar to the χ^2 difference test. The

fit index for a constrained model is compared to the unconstrained model, with a notable change indicating the presence of measurement bias. However, because sampling distributions do not exist for model fit indices, no critical value can be used to determine significance. Instead, practicing researchers have relied on recommendations from simulation research investigating $AFIs$ in measurement invariance testing. A few of the most commonly suggested $AFIs$ for examining measurement invariance include root mean square error of approximation ($RMSEA$; Steiger, 1989), comparative fit index (CFI; Bentler, 1990), Tucker-Lewis Index (TLI; Tucker & Lewis, 1973), and standardized root mean residual ($SRMR$).

1.3.3 Examining measurement invariance with $AFIs$

Alternative fit indices have been examined as possible detectors of measurement invariance to alleviate problems associated with the χ^2 difference test. Notably, Cheung and Rensvold (2002) examined the effect of the number of factors, number of items per factor, factor variance, correlations between factors, factor loadings, and sample size on model fit statistics when invariance is present. Results indicated that: 1) the $RMSEA$ is largely unaffected by the above study conditions, but showed larger standard errors in smaller samples, and 2) CFI and TLI decreased as the number of items and factors increased. The researchers concluded by suggesting $\Delta CFI < .01$ between measurement invariance conditions indicates the presence of invariance (Cheung & Rensvold, 2002).

Chen (2007) expanded on the work of Cheung and Rensvold (2002) by conducting two simulation studies examining the sensitivity of alternative fit indices when invariance was present and when noninvariant items existed at the weak, strong, and/or strict levels. When invariance was present between the two group, single factor model, the $SRMR$ showed more variability across the weak, strong, and strict models, than the CFI or $RMSEA$. In other words, the $SRMR$ was more sensitive to random sampling variability (Chen, 2007). In the second study, the effects of the proportion of invariance, number of indicators (i.e., 8 or 12), pattern of invariance (i.e., uniform with one group's model having higher loadings and intercepts, or non-uniform where loadings and intercepts varied in strength between the two groups), and ratio of sample size on alternative fit

indices for weak and strong tests of invariance were examined. A significant interaction between the pattern of invariance and proportion of invariance accounted for the most variability in all of the fit indices, including the CFI, $SRMR$, and $RMSEA$ when used to examine weak and strong invariance (Chen, 2007).

In particular, when the pattern of invariance was uniform, the examined fit indices showed the most change at 50% proportion of invariance, and the least change when the proportion was 0%, but when the pattern of invariance was mixed, the fit indices changed the least when the proportion of invariance was 0% and the most when the proportion was 75% (Chen, 2007). Furthermore, fit indices showed the most change during tests of weak and strong invariance when ratio of sample size was a balanced 1:1 ratio, suggesting the use of alternative fit indices for detecting a lack of weak or strong measurement invariance is best suited when groups sample sizes are balanced, and the proportion of invariance is both high and mixed across groups. Additionally, Meade, Johnson, and Braddy (2008) further examined the power, in other words, the ability to detect noninvariance (i.e., measurement bias) when it is present based on Cheung and Rensvold (2002) cutoff criteria. Meade et al. (2008) concluded that adequate power to detect noninvariance using the CFI and $RMSEA$ can be achieved for large sample sizes (N = 400 per group).

From these studies several rules-of-thumb and suggested cutoff criteria for change in alternative fit indices have been proposed when testing hypotheses of weak and strong measurement invariance. Cheung and Rensvold (2002) recommended $\Delta CFI < .01$ between the constrained model and the freely estimated model is sufficient for establishing weak or strong invariance across two groups, whereas both Chen (2007) and Meade et al. (2008) recommended more conservative values of $\Delta CFI < .005$, and $\Delta CFI < .002$, respectively. Furthermore, Chen (2007), recommended a $\Delta RMSEA \leq .01$ or .015 as a cutoff criterion for tests of both weak and strong invariance, and $\Delta SRMR \leq .025$ or .030 for weak invariance and $\Delta SRMR < .005$ or .010 for strong invariance. Conversely, although Meade et al. (2008) examined the $RMSEA$, the researchers found the variability of the $\Delta RMSEA$ across tests of weak and strong invariance to be too influenced by the number of items, factors, and sample sizes, and concluded that the $\Delta RMSEA$ should not be used when

evaluating measurement invariance. Interestingly, as an alternative to specific cutoff values, Little et al. (2007a) suggested the *RMSEA* for the constrained model be compared to the 90% confidence interval for the freely estimated model, with a *RMSEA* being within the interval suggesting that measurement invariance is established. However, an empirical examination of this technique was not provided.

1.3.4 Limitations of using CFA to evaluate measurement invariance

Multiple group CFA allows a researcher to evaluate the psychometric properties of a measure across populations through invariance testing, but some limitations exist. First, the method of identification used for the CFA model may influence tests of measurement invariance. For example, if a researcher uses the marker-variable method of identification, where the factor loading for one item across each group is fixed to 1.0, then the researcher is treating this item, known as the "referent" item, as invariant across groups. If this referent item is not invariant across groups, then tests of measurement invariance may not reveal this violation. If possible, referent items should be selected based on theory or previous research that supports invariance for the item (Reise et al., 1993), but this situation is not always possible. Previous researchers (Cheung & Rensvold, 1999; Cheung & Lau, 2012; Rensvold & Cheung, 2001) have addressed the above concern as a "standardization" problem and have proposed "factor-ratio test" as one potential solution.

The factor-ratio test involves the researcher repeatedly testing a desired level of measurement invariance between groups, with each new test having the researcher select a new referent item. In addition, one additional item, known as the "argument" is set constrained to be equal in the desired parameter across groups. A χ^2 difference test is conducted comparing this constrained model to a freely estimated model. This process continues until a constrained model with each item of a given factor has been selected as the referent and compared against each other item as an argument. An item is deemed invariant if the $\Delta\chi^2$ for that item as the argument was not significant across all possible referent items (Cheung & Rensvold, 1999). This test can be labor intensive, requiring $k(k-1)/2$ tests for each factor that consist of k items.

French and Finch (2008) evaluated power and Type I error rate for the factor-ratio test for testing weak invariance, concluding that although the procedure does maintain desired experimentwise α levels when a Bonferroni correction for multiple testing is used, the power of the procedure to detect noninvariant items decreased as model complexity (i.e., number of factors and items) and amount of noninvariant items increased. Unfortunately, researchers wishing to avoid the standardization problem by using the fixed-factor method of identification, where a factor variance is fixed to 1.0, may still encounter a problem. Yoon and Millsap (2007) state to avoid the fixed-factor method because if the factor variances are not equal across groups, then tests for weak invariance may result in a Type I error, where measurement bias may be found to exist simply due to improper equality constraint on the factor variance across groups.

Another limitation for using multiple group CFA to test measurement invariance may be how the alternative fit indices are calculated. For example, Widaman and Thompson (2003) have proposed that the null model used to calculate incremental fit indices, such as the CFI and TLI, in most structural equation modeling (SEM) software packages is incorrectly specified, and, in the case of longitudinal models, should be replaced by their suggested alternative null model. In SEM, the null model should be the worst fitting model to the data, where all item covariances are zero, and only means and variances are estimated for each item. When multiple groups and time points are present many software packages will default to estimating separate means and variances for the same item across group and time. When measurement invariance constraints across group and/or time are later imposed the constrained model may not be nested within the null model, and thus inappropriate for model comparisons involving χ^2 (Widaman & Thompson, 2003). Instead, Widaman and Thompson's (2003) proposed an alternative null model for the multiple group longitudinal CFA that should contain the following criteria:

1. Each item is a single indicator of its own latent variable
2. Variances for same latent variable across each time point and group are equated
3. Intercepts for the same item across each time point and group are equated

4. All item residuals are fixed to 0

5. All covariances among latent variables are fixed to 0

In conclusion, an examination of measurement invariance using multiple group CFA may require an alternative null model be specified before change in fit indices are examined for tests of measurement invariance.

1.4 Alternative Approaches for Examining Measurement Invariance

The examination of measurement invariance is not limited to a multiple group CFA framework. Indeed, multiple-indicators multiple-causes (MIMIC) models within the SEM framework have grown in popularity for examining a form of measurement invariance that is referred to as differential item functioning (DIF) in IRT literature. Similar to measurement invariance, DIF is an instance where an item on a scale performs differently across groups even when members of each group are matched on ability. If DIF exists, then a single item characteristic curve (ICC) cannot be used for both groups because the groups differ in either the estimated difficulty threshold (i.e., b parameter), the estimated discrimination (i.e., a parameter), or both parameters. Interestingly, there is a connection between a 2-PL IRT model and a categorical CFA where the a parameters are the same as factor loadings and the b parameters are the same as thresholds (see Raju et al., 2002).

Item parameter drift is one type of DIF that is also sometimes referred to as uniform DIF (De Ayala, 2009). In item parameter drift the difficulty changes across time due to perhaps test content being repeatedly taught or no longer attended too. Thus, the slope of the ICC remains the same, creating a uniform ICC, but the difficulty parameter changes. Item parameter drift would be most akin to having weak invariance established across time for a particular group of individuals, but a failure of strong invariance. One method for examining item parameter drift involves specifying a MIMIC model using the SEM framework.

Modern SEM software allows the researcher to easily specify a MIMIC model to examine group latent mean differences, as well as group differences in intercepts for each item. For example, two populations could be examined where a set of measured items are indicators for a latent construct. A dummy-coded variable for population membership could be examined as a predictor for each of the items. A significant slope from the dummy-coded group variable to a given item would indicate that there are differences in the item intercept between groups.

The advantage of the MIMIC model for testing item parameter drift is that is a much simpler and more parsimonious model than the multiple group CFA. Thus, model estimation and convergence rates will likely be much faster and higher, respectively. In addition, because only one model is specified, the researcher likely can perform tests of DIF with smaller samples than those required for multiple group CFA. The MIMIC model approach for testing DIF can also easily accommodate more than two groups by simply adding additional dummy codes. Recently, Woods and Grimm (2011) provided a method for testing non-uniform DIF where each item is regressed on a new latent variable that is specified as the interaction between the underlying latent construct and the grouping variable. If removal of the regression path between this new interaction latent variable and the item leads to significant change in model fit then there is evidence for possible non-uniform DIF (Woods & Grimm, 2011). Unfortunately, research on the best methods for estimating this latent variable interaction are still needed and current techniques implemented in software such as Mplus, can lead to increased Type I error rate for tests of DIF.

Indeed, a common disadvantage for using MIMIC models to examine DIF is that the method has been plagued with issues of multiple testing, including inflated Type I error rate (Kim et al., 2012b). Two suggested corrections for multiple testing issues in MIMIC models have been a Bonferroni correction or a different adjustment to the critical value in the likelihood ratio test as recommended by Oort (1998). Previous simulation studies by Kim et al. (2012b) have reported that the Bonferroni correction decreases power slightly, whereas the Oort adjustment both increases power when detecting only one DIF item, but it decreases power when two DIF items are present. Finally, Kim et al. (2012b) note the *CFI* and *SRMR* did not indicate model misspecification when

factor loadings, intercepts, or both contained "small" amounts of noninvariance (e.g., differences of 0.2).

1.5 Recent Advances in Examining Measurement Invariance in CFA

Previous research examining measurement invariance with CFA has largely focused on the simple two group case with a small number of factors and items. Recently, research has expanded to account for more complex models. For example, tests of measurement invariance assume that the items are linearly related to factor, but a situation may arise where an item has the same nonlinear relationship, such as a quadratic curve, with a factor for both groups. Tests of measurement invariance may fail, even though the same relationship exists across groups (Bauer, 2005). In a simulation study, Bauer (2005) notes that power to detect weak invariance quickly increases in the presence of moderate quadratic effects (i.e., < -0.050), suggesting that a failure of the test for weak invariance may be due to the nonlinear relationships and exploratory data analysis (e.g., scatter plots, & loess lines) should be employed to determine if a nonlinear relationship is present.

Another recent advance includes a demonstration of testing measurement invariance in CFA models that contain hierarchical, or second-order factors (Chen et al., 2005). The tests of configural, weak, strong, and strict invariance are conducted both on the lower order and higher order factors, with χ^2 difference tests, or comparisons of change in fit indices being examined for evidence of measurement invariance. However, Chen et al. (2005) note that earlier recommendations for change in fit indices (i.e.,Cheung & Rensvold, 2002) may or may not be warranted for examining measurement invariance in hierarchical models until future simulation research examines these claims.

Lastly, the discussion of measurement invariance in CFA has also moved to the examination of measurement invariance in multilevel models (Kim et al., 2012a). In a pair of simulation studies, Kim et al. (2012a) examined the effects of level 1 (i.e., within) and level 2 (i.e., between) unit

sizes, and intraclass correlation coefficients on the power of the χ^2 difference test, and the scaled χ^2 difference test (Satorra & Bentler, 2001) for detecting noninvariance in multiple group CFA and multilevel multiple group CFA models, respectively. Data were generated to reflect a treatment vs. control group situation where several level two units were present in both groups. A single factor model indicated by eight items was generated and noninvariance was defined as a 0.5 between-level factor loading difference on one of the eight items for the two different groups. Results revealed that ignoring a nested data structure and analyzing multilevel data with traditional multiple group CFA led to high Type I error rates, whereas multilevel CFA demonstrated lower Type I error rates and high power for detecting noninvariance across level 1 units (Kim et al., 2012a). Conversely, for detecting noninvariance across level 2 units, power of the scaled χ^2 difference test had a strong positive relationship with both the intraclass correlation coefficient and the number of level 2 units, suggesting that researchers wishing to have adequate power to detect weak invariance should strive to have an intraclass correlation coefficient > .33 and at least 80 level 2 units per each group (e.g., 160 level-2 units for a two group multilevel multiple group CFA; Kim et al., 2012a).

1.6 Summary

In summary, measurement invariance testing with factor analysis has a long history in social sciences and has particularly received increased attention over the last 30 years of research. Commonly, a series of nested model comparisons are conducted to examine if configural, weak, strong, and possibly, strict factorial invariance constraints are tenable across groups or time. When some, but not all of these model constraints hold, partial invariance is said to exist, but more research is needed to fully understand how much partial invariance is required before additional comparisons, such as tests of latent means, variances, and covariances, are warranted. Tests of measurement invariance can involve model comparisons via a χ^2 difference test, or change popular model fit indices including *RMSEA*, *CFI*, *TLI*, and *SRMR*. Although previous research has made notable progress on evaluating measurement invariance based on change in model fit indices, future re-

search is needed to examine how these indices are influenced by increased model complexity, including multiple group longitudinal research designs. Lastly, future research should examine the use of alternative null models, and unbalanced sample sizes in invariance testing.

1.7 The Current Study

Monte Carlo simulation studies are common in structural equation modeling when the researcher wishes to examine properties of statistics, such as model fit indices (Bandalos, 2006; Bandalos & Gange, 2012). Although previous simulation studies (Chen, 2007; Cheung & Rensvold, 2002; Meade et al., 2008) were thorough investigations of measurement invariance, none of these studies examined how well alternative fit indices function in detecting measurement invariance in longitudinal models or models with more than two groups. Moreover, Widaman and Thompson (2003) have proposed that the null model used to calculate fit indices, including the CFI and TLI, in most SEM software packages is incorrectly specified, and, in the case of longitudinal models, should be replaced by their suggested alternative null model. The current study expands on previous research in measurement invariance by further examining the power of $AFIs$, including the CFI, TLI, $RMSEA$, and $SRMR$, for longitudinal multiple group models when varying amounts of noninvariance is present and an appropriately specified alternative null model is estimated.

1.7.1 Hypotheses

A Monte Carlo simulation study was conducted with the following study characteristics manipulated: type of mixed design, sample size, sample size ratio, type of noninvariance (i.e., weak, strong), location of invariance (i.e., group or time), and amount of weak/strong noninvariance (i.e., measurement bias effect size). Specific hypotheses for the current study are described below.

1.7.1.1 Hypothesis 1

Previous research has noted that power for tests of measurement invariance decreases as models become more complex, including more items (Chen, 2007; Meade et al., 2008) or factors (Meade et al., 2008). Power to detect weak or strong measurement bias was hypothesized to be lower for more complex (i.e., more groups and time points) mixed designs.

1.7.1.2 Hypothesis 2

Chen (2007) reported larger $\Delta AFIs$ for tests of invariance when sample sizes were balanced between groups than unbalanced. Greater $\Delta AFIs$ for balanced samples implies more power to detect noninvariance in these conditions because larger $\Delta AFIs$ are more likely to exceed a predetermined cut-off values. Therefore, power to detect weak or strong measurement bias was hypothesized to decrease as the sample size ratio increases.

Chapter 2

Methods

2.1 Data Generation

The following study consisted of two phases. In Phase 1, models with various mixed designs, sample sizes, and sample size ratios (see Study Conditions below) were generated with strong invariance present across both groups and time. For each replication, the generated model fit statistics (see Measures below) were recorded for tests of weak, and strong invariance. The change in fit index from configural to weak invariance and from weak to strong invariance was recorded to create a sampling distribution of the change in fit index (i.e., ΔAFI) due to random sampling variability. The 95th percentiles for the $\Delta RMSEA$, and $\Delta SRMR$ within each test of invariance and within each study condition were selected as cut-off values for Phase 2. In addition, the 5th percentiles for ΔCFI, and ΔTLI calculated using the software default null model, and for ΔCFI_A and ΔTLI_A calculated using the alternative null model (see Widaman & Thompson, 2003) from both the tests of weak and strong invariance were selected as a cut-off values for Phase 2.

Recall, in Phase 1 strong invariance was present. Specifically, across group and time for each latent variable the factor loadings (λ) were fixed to 0.7, intercepts (τ) were fixed to 0 and residuals (θ) were fixed to 0.51 in the population model. In addition, factor variances were fixed to 1.0, and factor covariances (Ψ), which become factor correlations because the factor variances were

fixed to 1.0, between adjacent time points were fixed to 0.3 and were specified to follow a perfect simplex structure (see Guttman, 1954; Jöreskog, 1970; Little, 2013) when more than two time points were present. For example, $\Psi_{Time_2,Time_1} = \Psi_{Time_3,Time_2} = 0.3$, and $\Psi_{Time_3,Time_1} = 0.3 * 0.3 = 0.09$. Similarly, because longitudinal data were simulated, the covariances for residuals of same indicator were specified to equal 0.3 across each adjacent time point and to also follow a perfect simplex structure (e.g., $\Theta_{4,1} = \Theta_{7,4} = 0.3$ and $\Theta_{7,1} = 0.09$).

The above model parameters were selected to mimic results reported in applied multiple group and/or longitudinal tests of measurement invariance (e.g., Atienza et al., 2003; Barbosa-Leiker et al., 2013; Bowers et al., 2010; Pitts et al., 1996; Short & Hawley, 2012; Wu et al., 2009). Furthermore, these model parameters were similar to those used in previous methodological examinations of measurement invariance testing (Chen, 2007; Cheung & Rensvold, 2002; Meade et al., 2008).

All models were identified using the fixed factor method of identification by fixing the latent variable's variance to 1.0. Specifically, for tests of weak invariance, the factor variance at time 1 and in group 1 was fixed to 1.0. The remaining factor variances across each group and time point were freed, as the set scale for the construct at time 1 in group 1 and the factor loadings constrained to equality across both group and time is sufficient for identifying the model (see Little, 2013). Likewise, the means structure must be identified for tests of strong invariance. The latent mean at time 1 and in group 1was fixed to 0 (see Hancock, 2004 for a detailed justification of identifying the means structure) to set the scale. The remaining latent means across each group and time point were freely estimated as the scale was set in the construct for group 1 at time 1 and the item intercepts constrained to equality across group and time were again sufficient in identifying the model.

2.2 Study Conditions

Simulation studies are typically prone to criticism for their generalizability to real data. The below study conditions were chosen to reflect a moderately well-fitting model that could be possible in

real data. Moreover, these specific parameter values have been used in previous simulation studies examining power in multiple group CFA (see Hancock et al., 2000; Hancock, 2001; Meade et al., 2008). In addition, smaller total sample sizes (e.g., N = 300) and time points (e.g., 2) were chosen as levels for some conditions to reflect data an applied researcher may have obtained or have the ability to collect. The present study was a 2 (mixed design type) x 2 (location of noninvariance) x 2 (lack of invariance) x 10 (amount of noninvariance) x 4 (sample size) x 3 (sample size ratio) x 2 (null model type) design.

2.2.1 Mixed design type

Previous examinations of *AFIs* in measurement invariance testing (e.g., Chen, 2007; Cheung & Rensvold, 2002; Meade et al., 2008) have examined invariance testing in only two groups at a single time point. The current study expanded on the previous research by examining invariance testing in two common mixed factorial designs, including the 2 (group) x 2 (time), and 3 (group) x 3 (time) design. For example, the 3 (group) x 3 (time) mixed design condition consisted of three groups repeatedly measured across three time points.

2.2.2 Location of noninvariance

The location of noninvariance consisted of two conditions: noninvariance between groups or noninvariance across time. Specific details about the location of noninvariance conditions are described below.

2.2.3 Amount of noninvariance

Two different levels were generated for lack of invariance, including lack of weak invariance (i.e., unequal factor loadings) and lack of strong invariance (i.e., unequal item intercepts). Lack of invariance was specified as Group 2 differing from Group 1 in the 2 (group) x 2 (time) design, and Group 3 differing from Groups 1 and 2 in the 3 (group) x 3 (time) design. Thus, Group 1 in

the 2 (group) x 2 (time) design, and Groups 1 and 2 in the 3 (group) x 3 (time) design could be considered the "reference" groups. Group 2 in the 2 (group) x 2 (time) design, and Group 3 in the 3 (group) x 3 (time) design could be considered the "focal" groups. For both weak and strong lack of invariance, one of the three items indicating the single factor (i.e., 33% of the items) in the focal group differed in either factor loadings for lack of weak invariance or item intercepts for lack of strong invariance. The specified amounts of weak and strong noninvariance are described below.

Figures 2.1 and 2.2 display the population models with lack of weak (color coded in blue) and strong (color coded in red) invariance across time for the 2 (group) x 2 (time) and 3 (group) x 3 (time) designs, respectively. Figures 2.3 and 2.4 display the population models with lack of weak (color coded in blue) and strong (color coded in red) invariance between groups for the 2 (group) x 2 (time) and 3 (group) x 3 (time) designs, respectively

2.2.3.1 Weak noninvariance effect sizes

Because there were two locations of noninvariance (i.e, group or time) the amount of weak noninvariance was manipulated in each condition. In the conditions of weak noninvariance between groups, item 3 in the focal group was specified to have factor loadings decrease ($\lambda - \Delta\lambda$) ranging from 0 (i.e., invariance) to 0.4 in units of .04 for all time points. In other words, the factor loading for item 3 could be 0.7, 0.66, 0.62, 0.58, 0.54, 0.5, 0.46, 0.42, 0.38, 0.34, and 0.30.

In the lack of weak invariance across time conditions, the factor loading for item 3 were specified to decrease from 0 to 0.4 in units of .04 from Time 1. This factor loading decrease across time was the same for each group to represent a lack of weak invariance across time, but the presence of weak invariance across group.

2.2.3.2 Strong noninvariance effect sizes

Similar to the weak noninvariance effect sizes, the amount of of strong noninvariance was manipulated between groups and across time. In the lack of strong invariance between between groups condition, the intercept (τ) for item 3 increased ($\Delta\tau$) from 0 to 0.4 in units of .04 for the focal

Figure 2.1: Population Model for 2 (group) x 2 (time) Condition with Lack of Weak or Strong Invariance across Time

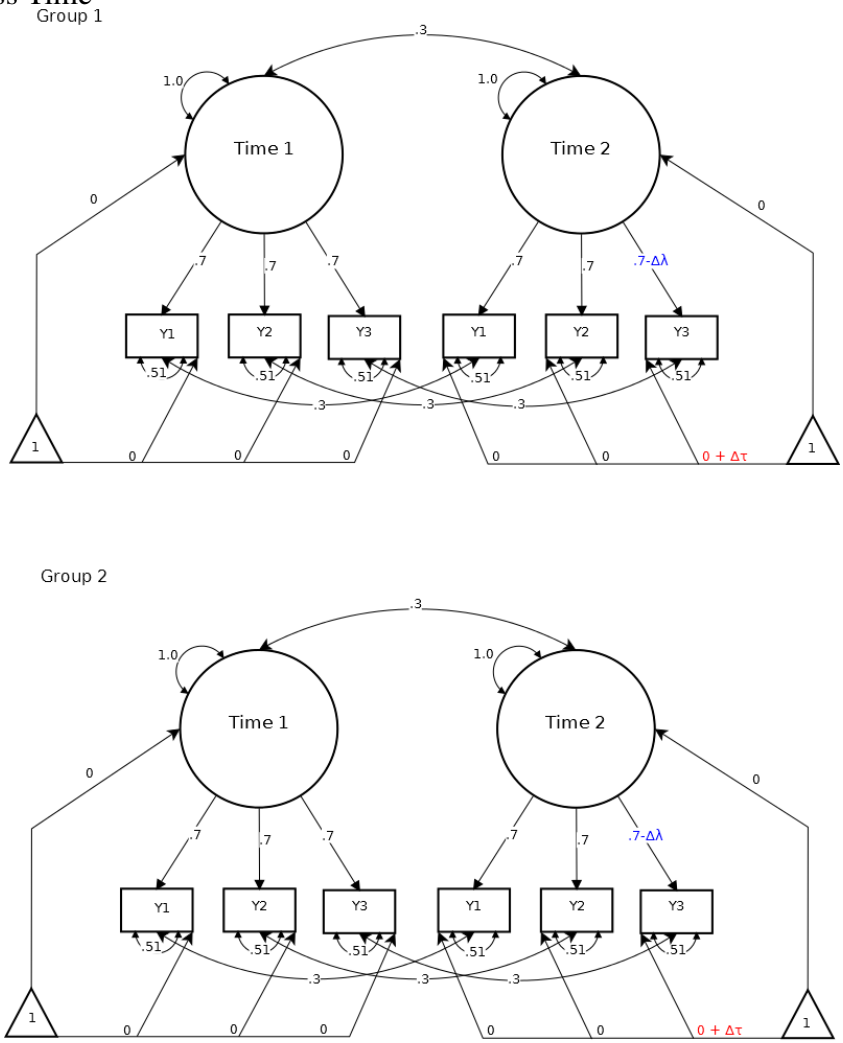


Figure 2.2: Population Model for 3 (group) x 3 (time) Condition with Lack of Weak or Strong Invariance across Time

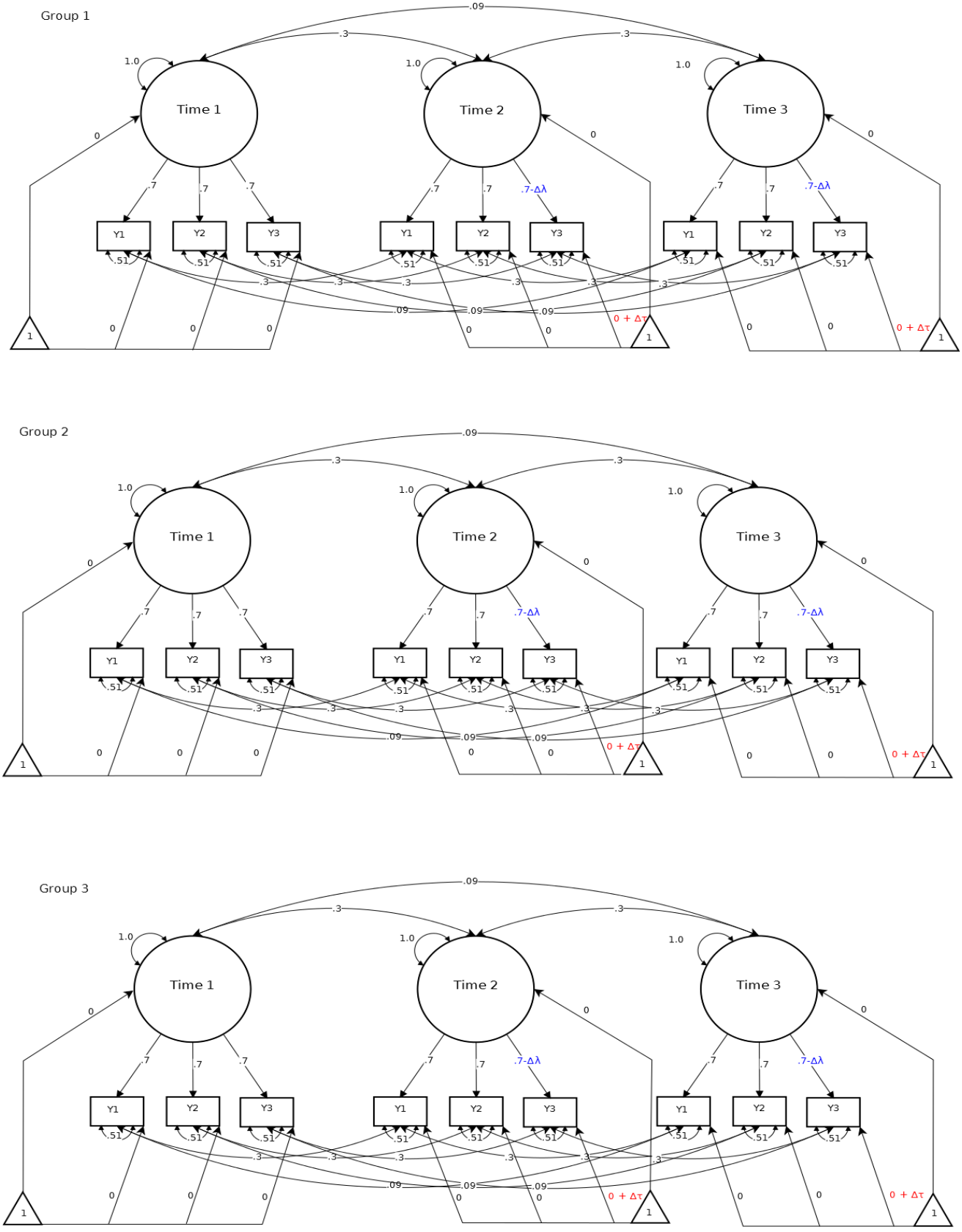


Figure 2.3: Population Model for 2 (group) x 2 (time) Condition with Lack of Weak or Strong Invariance between Groups

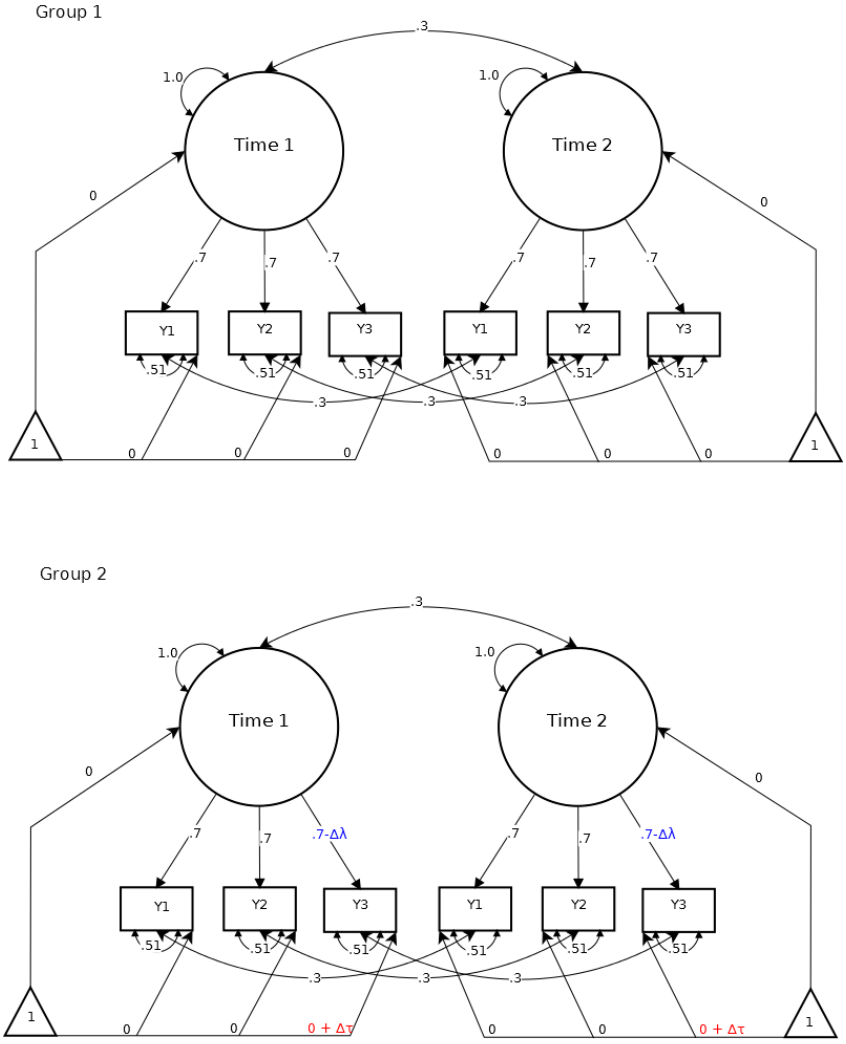
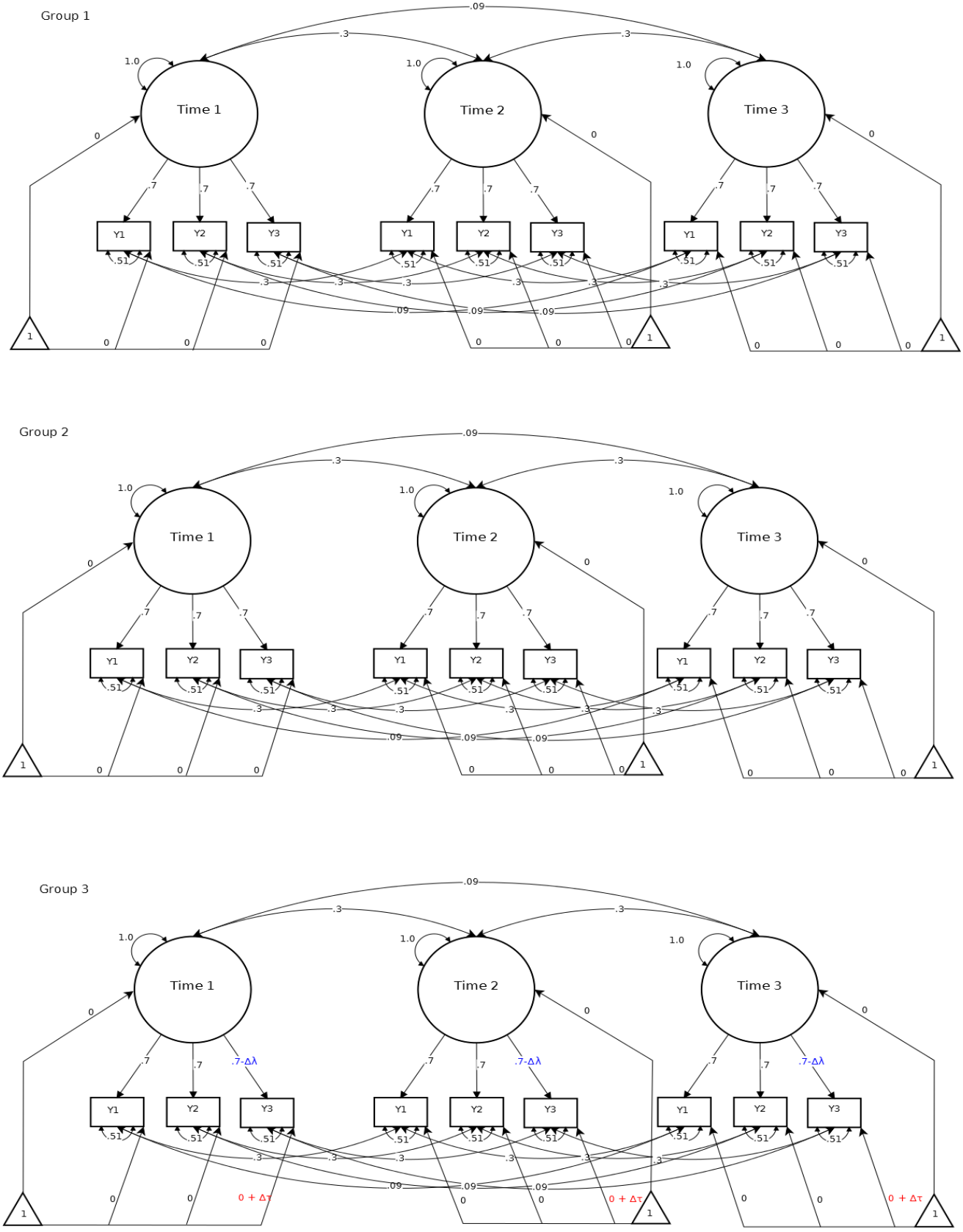


Figure 2.4: Population Model for 3 (group) x 3 (time) Condition with Lack of Weak or Strong Invariance between Groups



group across all time points. Lastly, for the lack of strong invariance across time condition, item 3 increased from 0 to 0.4 by units of .04 for each time point following Time 1 in each group.

2.2.4 Total sample size

Four conditions of sample size were created for both the 2 (group) x 2 (time) design and the 3 (group) x 3 (time). Specifically, for the 2 x 2 design total sample sizes were N = 300, 600, 900 and 1200, and for the 3 x 3 design the sample sizes were 360, 720, 1080, and 1800.

2.2.5 Sample size ratio

Three conditions were created to examine the effects of having unbalanced sample sizes in a study. Sample size ratios in the 2 (group) x 2 (time) design were 1:1, 2:1, and 4:1, and in the 3 (group) x 3 (time) design were 1:1:1, 2:1:1, and 4:1:1. Table 2.1 provides the sample size ratios for each possible sample size and group condition

Table 2.1: Sample Size Ratio by Mixed Design Type and Total Sample Size

2 (group) x 2 (time) Design				
	Total Sample Size (N)			
Sample Size Ratio	300	600	900	1200
1:1	150:150	300:300	450:450	600:600
2:1	200:100	400:200	600:300	800:400
4:1	240:60	480:120	720:180	960:240
3 (group) x 3 (time) Design				
	Total Sample Size (N)			
Sample Size Ratio	360	720	1080	1800
1:1:1	120:120:120	240:240:240	360:360:360	600:600:600
2:1:1	180:90:90	360:180:180	540:270:270	900:450:450
4:1:1	240:60:60	480:120:120	720:180:180	1200:300:300

2.2.6 Null model

Following the recommended alternative null model guidelines from Widaman and Thompson (2003) both the alternative null model and software default null model were estimated and used

in the calculation of the incremental fit indices, specifically the CFI and TLI, for all models and across all conditions. Fit statistics calculated with the alternative null model were denoted with the subscript $_A$ (e.g. CFI_A).

2.3 Procedure

All analyses were conducted on a high performance computer cluster that consisted of Dell PowerEdge 2950 and 1950 systems. First, multivariate normal data sets were generated based on covariance matrices and mean structures derived from the population models described above using the *mvtnorm* package version 0.9-9996 (Genz & Bretz, 2009) in the software R version 3.0.1 (R Core Team, 2013). Next, each data set was fit to configural, weak, and strong invariant models using the *lavaan* package version 0.5-16 (Rosseel, 2012) in R. Each model was fit using maximum likelihood estimation and was given 10,000 iterations to converge. The *lavaan* default for start values was used for each model. Specifically, start values for factor means, factor covariances item intercepts, and residual covariances were set to zero. Start values for factor loadings were set to one, and residual start values were fixed to half of the observed variance. These start values are similar to those in other popular software packages such as *Mplus* (Muthén & Muthén, 2010). Fit statistics (see Measures below) for each of these models were recorded and the change in these fit indices (i.e, $\Delta AFIs$) from the configural to the weak, and from the weak to the strong invariant model were recorded and used to calculate power (see Measures below). Incremental fit statistics (i.e., CFI & TLI) and change in incremental fit statistics (i.e., ΔCFI & ΔTLI) were calculated with both the software default independence null model and Widaman and Thompson's (2003) recommended alternative null model for each analyzed data set.

2.4 Measures

2.4.1 Chi-square (χ^2)

Traditionally, model fit in CFA has been evaluated by examining the χ^2 from the estimated model, which compares the model implied covariance matrix to the observed covariance matrix. The χ^2 for each model was measured, and the $\Delta\chi^2$ was recorded for each test of measurement invariance.

2.4.2 Root mean square error of approximation (*RMSEA*)

The *RMSEA* (Steiger, 1989) is an absolute fit index, meaning the estimated model is compared to a saturated model with perfect fit. The *RMSEA* provides an estimate of how much error per degree of freedom a model has and can be calculated as follows:

$$RMSEA = \sqrt{\frac{\left[\frac{\left(\chi_t^2 - df_t\right)}{(N-1)}\right]}{\left(\frac{df_t}{g}\right)}} \tag{2.1}$$

Where χ_t^2 = the tested (i.e., estimated) model's χ^2, df_t = the tested model's degrees of freedom, N = the sample size, and g =the number of groups. The *RMSEA* and $\Delta RMSEA$ were recorded for each test of measurement invariance.

2.4.3 Comparative fit index (*CFI*)

The *CFI* (Bentler, 1990) is an incremental fit index, where the estimated model is compared back to a null model. The CFI provides a ratio of misfit from the tested model compared to the null model

$$CFI = \frac{max\left[\left(\chi_t^2 - df_t\right), 0\right]}{max\left[\left(\chi_t^2 - df_t\right), \left(\chi_0^2 - df_0\right), 0\right]} \tag{2.2}$$

where χ_t^2 = the tested model's χ^2, df_t =the degrees of freedom for the tested model, χ_0^2 = the null model χ^2, df_0 =the null model degrees of freedom. The *CFI* and ΔCFI, as well as the

CFI_A and ΔCFI_A using the alternative null model, were recorded for each test of measurement invariance.

2.4.4 Tucker-Lewis index (TLI)

The TLI (Tucker & Lewis, 1973) is another incremental fit index that indicates the ratio of model misfit for tested model, subtracted from the ratio of misfit from the null model that is then divided by the ratio of misfit from the tested model minus one.

$$TLI = \left[\frac{\left(\frac{\chi_0^2}{df_0}\right) - \left(\frac{\chi_t^2}{df_t}\right)}{\left(\left(\frac{\chi_t^2}{df_t}\right) - 1\right)} \right] \tag{2.3}$$

where $\chi_t^2 =$ the tested model's χ^2, $df_t =$the degrees of freedom for the tested model, $\chi_0^2 =$ the null model χ^2, $df_0 =$the null model degrees of freedom. The TLI and ΔTLI, as well as the TLI and ΔTLI_A using the alternative null model, were recorded for each test of measurement invariance.

2.4.5 Standardized root mean residual (*SRMR*)

The *SRMR* is an absolute fit index that examines model misfit based on the residuals by providing a standardized average amount of misfit per observed variable (i.e., indicator). The *SRMR* can be calculated by the following formula:

$$SRMR = \sqrt{\frac{\left\{ \sum_{i=1}^{p} \sum_{j=1}^{i} \left[\frac{\left(s_{ij} - \hat{\sigma_{ij}}\right)}{\left(s_{ii} s_{jj}\right)} \right]^2 \right\}}{p(p+1)}} \tag{2.4}$$

where $p =$ the number of observed variables, s_{ii} and s_{jj} are the observed standard deviations, $s_{ij} =$ observed covariances, and $\hat{\sigma}_{ij} =$estimated covariances.The *SRMR* and $\Delta SRMR$ were recorded for each test of measurement invariance.

2.4.6 Akaike Information Criterion (*AIC*)

The Akaike information criterion (AIC; Akaike, 1987) accounts for how well the model fits the data with a penalty added for model complexity (see Burnham & Anderson, 2004 for more details). When comparing two models, the model with a lower *AIC* is considered more replicable and should be retained. The *AIC* has traditionally been suggested for comparisons of models that are not nested (Brown, 2006; Kline, 2011) however, this measure of fit has been noted for possible use in invariance testing (Little et al., 2007b; van de Schoot et al., 2012). Interestingly, sensitivity of the *AIC* in the configural model was reported by Cheung and Rensvold (2002), but the use of this fit measure for tests of weak or strong invariance was not examined by later researchers (e.g., Chen, 2007; Meade et al., 2008). Furthermore, the *AIC* has been used alongside $\Delta AFIs$ in substantive investigations of measurement invariance using CFA (e.g., Wicherts et al., 2004; Wicherts & Dolan, 2010) and with exploratory structural equation modeling (ESEM; Marsh et al., 2009). Thus, the use of *AIC* for tests of multiple group longitudinal tests of measurement invariance is worth further investigation and was included in the current study.

Several calculations for the *AIC* exist (see Brown, 2006), including:

$$AIC = \chi^2 - 2df \tag{2.5}$$

$$AIC = \chi^2 - 2p \tag{2.6}$$

$$AIC = -2(loglikelihood) + 2p \tag{2.7}$$

where χ^2 = the tested model's χ^2, df =the degrees of freedom for the tested model, and p = the estimated number of parameters in the model. Kaplan (2009) describes how each of these above equations are equivalent. Equation 2.5 was used to calculate *AIC* in the current study. Wicherts and Dolan (2004) note the in the case of examining measurement invariance it is important to have

each nested model include the means structure, even if it is saturated, in order to make proper model comparisons using the *AIC*. Thus, in the current study, the mean structure was included in the configural, weak, and strong models when calculating *AIC*.

2.4.7 Power

Power is the probability of rejecting the null hypothesis, given that the null hypothesis is false. In order to determine power of the $AFIs$, excluding *AIC*, for detecting noninvariance, a cut-off value for the fit indices was established (see Phase 1 of Data Generation above). The number of replications within a condition with a specified lack of invariance that exceeded the condition specific cut-off value established in Phase 1 of the data generation were tallied to reflect power. Power of *AIC* for tests of measurement invariance was calculated as the number of replications within each condition where *AIC* for the constrained model was less than the *AIC* for the unconstrained model.

Chapter 3

Results

3.1 Model Convergence and Improper Solutions

Within each condition 0-10.6% of the replications failed to converged. Figures A.1 and A.2 in Appendix A display the rates of non-converged solutions for each study condition during tests of weak and strong invariance, respectively. A closer examination across each condition revealed that non-convergence occurred in conditions when the samples were unbalanced (e.g., sample size ratio of 4:1 or 4:1:1) and total sample sizes were N = 300 in the 2 (group) x 2 (time) design, and N = 360 in the 3 (group) x 3 (time) design. A sample of 10 data sets from the 3 (group) x 3 (time) condition where $\Delta\lambda = 0.4$, N = 360, and sample size ratio = 4:1:1 that failed to converge using *lavaan* were extracted and reanalyzed using the popular software package *Mplus* version 7 (Muthén & Muthén, 2010). None of these data sets converged to the configural model after 10,000 iterations. Thus, choice of SEM software was not considered a problem for non-convergence. Conversely, only 0-0.4% of replications did not converge in tests on strong invariance. In other words, non-convergence was much more common during tests of weak invariance. Non-converged replications were removed from further analyses.

Improper (i.e., inadmissible) solutions, were quite common in certain conditions. Figures A.3 and A.4 display the percentage of improper solutions for tests of weak invariance and strong invari-

ance, respectively. Tests of weak invariance contained between 0-76% inadmissible solutions. The rate of improper solutions was associated with sample size, sample size ratio, and amount of weak noninvariance (i.e., change in factor loading) for tests of weak invariance. The rate of improper solutions increased as the change in factor loading increased, and this effect appeared greater when total sample size decreased and sample sizes among groups became more unbalanced. Tests of strong invariance contained between 0-14% improper solutions. In particular, improper solutions occurred across 14% of the replications in the 3 (group) x 3 (time) design when the total sample size was N = 360 and the sample size ratio was 4:1:1. In the remaining study conditions the percentage of improper solutions for tests of strong invariance ranged between 0-2%.

Further investigation among the tests of weak and strong invariance revealed that improper solutions only occurred in the configural invariance model. The improper solutions were due to the presence of a nonpositive definite residual matrix in the configural invariance model. Specifically, the focal group contained a negative residual variance (i.e. Heywood case) or a residual correlation > 1.0 in > 95% of the occasions when an improper solution occurred. In the remaining < 5 % of occasions, group 2 in the 3 (group) x 3 (time) design contained a Heywood case or a residual correlation > 1.0 in configural invariance model.

The presence of improper solutions was only a concern for tests of weak invariance, because a test of weak invariance is a nested model comparison between the weak invariance model and the configural invariance model. Improper solutions were not a concern for tests of strong invariance because the test of strong invariance was a nested model comparison between the strong invariance model and the weak invariance model, and improper solutions did not occur in either of these models.

How to properly deal with improper solutions in analyses has been of topic of interest for CFA methodologists for several decades. Researchers may be inclined to constrain a Heywood case to a particular value, such as 0, but this tactic is discouraged because evaluating model χ^2 test statistic becomes less straightforward. The χ^2 statistic may no longer follow a chi-square distribution, but instead of mixture of chi-square distributions, where properties of this mixture distribution are

likely unknown to the researcher (Savalei & Kolenikov, 2008). Instead, Savalei and Kolenikov (2008) suggest researchers do not constrain a Heywood case to 0, but instead use its presence as evidence of possible model misspecification. Interestingly, improper solutions are more likely to occur in studies with smaller samples (Anderson & Gerbing, 1984; Chen et al., 2001; Gerbing & Anderson, 1987) and factor loadings (Anderson & Gerbing, 1984; Gerbing & Anderson, 1987). Previous measurement invariance research by Meade et al. (2008) also reported that a majority of their analyses contained inadmissible solutions, where the proportion of inadmissible solutions was strongly related to model misspecification (e.g., the frequency of inadmissible solutions increased as the number of noninvariant items increased).

Model fit, specifically change in model fit (i.e., $\Delta AFIs$) was the primary focus of this study. Thus, the primary concern was if model fit statistics from improper solutions differed from proper solutions. Gerbing and Anderson (1987) reported the χ^2, and the goodness of fit index (GFI) did not significantly differ between models with proper and improper solutions. Furthermore, although the root mean residual (RMR) was significantly higher in improper solutions than proper solutions, the small difference was noted as not practically significant (Gerbing & Anderson, 1987). Chen et al. (2001) reported similar results, noting no practical differences between χ^2 statistics from models with proper solutions and models with improper solutions. Previous examinations by Meade et al. (2008) have reported the RMR to be highly related to $SRMR$ ($r = .99$), TLI ($r = . - .90$) and CFI ($r = -.87$), as well as the GFI to be highly correlated with CFI ($r = .90$) and TLI ($r = .86$). Given the recommendations for handling negative error variances (Savalei & Kolenikov, 2008) and evidence from previous research (Anderson & Gerbing, 1984; Chen et al., 2001; Gerbing & Anderson, 1987; Meade et al., 2008), $AFIs$ from improper solutions in the current study were kept and used in further analyses described below.

3.2 ΔAFI Cut-off Values

During Phase 1 of the study, 500 replications of strong invariant models were generated for each of the study conditions (i.e., design type, total sample size, sample size ratio). Weak and strong invariance tests were performed on these models and the $\Delta AFIs$ were recorded for both tests. Because strong invariance was present in the data generation models, any $\Delta AFIs$ were due to sampling variability. Thus, the $\Delta AFIs$ for tests of weak and strong invariance in each unique study condition were treated as sampling distributions and the 95$^{\text{th}}$ percentiles for $\Delta RMSEA$ and $\Delta SRMR$, and the 5$^{\text{th}}$ percentiles for ΔCFI, ΔCFI_A, ΔTLI, and ΔTLI_A were recorded as cut-off values (e.g., critical values for $\alpha = .05$). Appendix B contains tables with the specific calculated cut-off values for each study condition.

3.3 Relationships Among $\Delta\chi^2$ and $\Delta AFIs$

Table 3.1 displays the correlations among $\Delta\chi^2$ and $\Delta AFIs$ across all study conditions for tests of weak invariance (e.g., $AFI_{weak} - AFI_{configural}$) below the diagonal, and tests of strong invariance (e.g., $AFI_{strong} - AFI_{weak}$) above the diagonal. Overall, the $\Delta AFIs$ demonstrated strong correlations among each other in the expected patterns. When model constraints during tests of invariance are imposed, model fit will be worse, resulting in an increase χ^2, $RMSEA$, and $SRMR$, and decrease in both types of CFI and TLI. The use of an alternative null model had little effect on the ΔCFI and ΔTLI, with the correlations between ΔCFI and ΔCFI_A, and ΔTLI and ΔTLI_A both exceeding .99 for tests of weak and strong invariance. The $\Delta SRMR$ had the lowest relationships with any of the other $\Delta AFIs$.

3.4 Influence of Study Conditions on $\Delta\chi^2$ and $\Delta AFIs$

The influence of study conditions on the change in χ^2 and $AFIs$ was examined for tests of both weak and strong invariance by conducting a 2 (mixed design type) x 4 (sample size) x 3 (sample

Table 3.1: Correlations among $\Delta AFIs$ for Tests of Weak and Strong Invariance

Measure	$\Delta\chi^2$	$\Delta RMSEA$	ΔCFI	ΔCFI_A	ΔTLI	ΔTLI_A	$\Delta SRMR$
$\Delta\chi^2$	—	0.757	-.761	-.762	-.765	-.748	.749
$\Delta RMSEA$	.675	—	-.915	-.917	-.948	-.947	.911
ΔCFI	-.717	-.868	—	.999	.973	.973	-.876
ΔCFI_A	-.716	-.868	.999	—	.974	.973	-.876
ΔTLI	-.693	-.916	.948	.948	—	.999	-.898
ΔTLI_A	-.667	-.916	.946	.946	.998	—	-.894
$\Delta SRMR$	.516	.784	-.782	-.782	-.816	-.822	—

Note. Correlations among change in fit indices for tests of weak invariance are below the diagonal, whereas correlations among change in fit indices for tests of strong invariance are above the diagonal. Because the amount of change in *AIC* is not evaluated in model comparisons, ΔAIC is not included in the above table.

size ratio) x 10 (weak noninvariance effect size) x 2 (location of noninvariance) ANOVA for $\Delta\chi^2$ and each ΔAFI. Results from each ANOVA are discussed below. Predictors that accounted for > 3% of the variability in the $\Delta\chi^2$ or $\Delta AFIs$ are discussed below[1].

3.4.1 Tests of weak invariance

Table 3.2 displays the amount of variance explained (η^2) for tests of weak invariance. Weak noninvariance effect size accounted for a majority of the variability (39.39 - 60.85%) in $\Delta\chi^2$ and each of the $\Delta AFIs$. The $\Delta\chi^2$ and $\Delta SRMR$ were also influenced by the design type (6.93% and 6.83%, respectively), with larger change in these fit indices being observed in the 3 (group) x 3 (time) design. A Sample Size x Weak Noninvariance Effect Size accounted for 11.06% of variability in the $\Delta\chi^2$, where the increase in weak noninvariance led to larger $\Delta\chi^2$ and this increase was greater in lager samples. As expected, sample size also accounted for 13.78% of the variability in $\Delta\chi^2$, with larger samples sizes leading to larger $\Delta\chi^2$.

In addition, 3.36 - 3.92% of variability in ΔCFI, ΔCFI_A, ΔTLI, and ΔTLI_A was accounted for by a Weak Noninvarinace Effect Size x Location of Noninvariance interaction, where the increase in weak noninvariance led to larger change in these fit indices and this increase was greater when

[1]Chen (2007) reported effects that accounted for > 2% of the variability in the change of the fit index. Given the large number of study conditions, effects accounting for > 3% in the current study were chosen to aid interpretation.

Table 3.2: Percent Variance Explained by Study Conditions on the Change in Fit Indices for Tests of Weak Invariance

Predictors	$\Delta\chi^2$	$\Delta RMSEA$	ΔCFI	ΔCFI_A	ΔTLI	ΔTLI_A	$\Delta SRMR$
Sample Size (*N*)	**13.78**	1.17			0.02	0.02	0.44
Sample Size Ratio (R)	0.42	0.48	0.70	0.70	0.71	0.70	1.89
Weak Noninvariance Effect Size (WN)	**39.39**	**56.22**	**60.83**	**60.85**	**59.53**	**58.03**	**46.10**
Location of Noninvariance (LN)	1.99	1.78	**4.57**	**4.57**	**4.32**	**4.10**	0.53
Design Type (DT)	**6.93**	2.65	0.89	0.89	0.98	1.93	**6.83**
N x R	0.14		0.01	0.01			
N x WN	**11.06**	0.34	0.06	0.06			0.86
R x WN	0.23	0.20	0.60	0.60	0.52	0.52	0.40
N x LN	0.63						0.01
R x LN	0.49	0.59	0.85	0.85	0.81	0.80	1.69
WN x LN	1.29	0.70	**3.92**	**3.92**	**3.51**	**3.36**	0.46
N x DT	0.19	0.19	0.01	0.01			0.02
R x DT		0.03	0.07	0.07	0.07	0.09	0.20
WN x DT	0.71	1.59	0.81	0.82	0.88	1.66	2.68
LN x DT	0.16		0.06	0.06	0.05	0.01	0.10
N x R x WN	0.06						0.01
N x R x LN	0.14						0.03
N x WN x LN	0.39						0.01
R x WN x LN	0.25	0.24	0.67	0.68	0.60	0.61	0.91
N x R x DT							
N x WN x DT	0.16	0.04	0.02	0.02	0.00	0.00	0.10
R x WN x DT		0.01	0.04	0.04	0.03	0.05	0.04
N x LN x DT	0.05						0.02
R x LN x DT		0.02	0.06	0.06	0.05	0.08	0.20
WN x LN x DT	0.11		0.09	0.09	0.07	0.04	0.09
N x R x WN x LN	0.06						0.01
N x R x WN x DT							
N x R x LN x DT							
N x WN x LN x DT	0.03						
R x WN x LN x DT			0.03	0.03	0.03	0.04	0.04
N x R x WN x LN x DT							

Note. Percent variance explained estimates are η^2 effect sizes. Model effects > 3% are in bold. Model effects that accounted for percent of variance explained < 0.01 % are removed. Because the amount of change in *AIC* is not evaluated in model comparisons, ΔAIC is not included in the above table.

noninvariance was present across time versus group. The study conditions did not seem to differentially influence $\Delta AFIs$ calculated with the software default versus the suggested alternative null model. In fact, η^2 effect sizes reported for the ΔCFI and ΔTLI were quite similar to the same fit statistics calculated with the suggested alternative null model.

3.4.2 Tests of strong invariance

Table 3.3 displays the amount of variance explained (η^2) for tests of strong invariance. Similar to the test of weak invariance results, the amount of noninvariance (i.e., strong noninvariance effect size) accounted for most (35.27-60.18%) of the variability in $\Delta\chi^2$ and each of the $\Delta AFIs$. Furthermore, the location of noninvariance accounted for 6.26 - 20.70% of the variability in $\Delta\chi^2$ and each of the $\Delta AFIs$, with greater $\Delta AFIs$ being observed for noninvariance across time versus group. In addition, the design type influenced each of the $\Delta AFIs$ (3.05-4.74%) with larger change in these fit indices being observed in the 3 (group) x 3 (time) design.

A Sample Size x Strong Noninvariance x Location of Noninvariance interaction accounted for variability (3.35%) in $\Delta\chi^2$. Probing this interaction revealed that the Sample Size x Strong Noninvariance Effect Size interaction, where a positive effect of sample size was greater as the amount of noninvariance increased, was more pronounced when the location of noninvariance was between groups than across time. Again, sample size also accounted for variability (11.13%) in $\Delta\chi^2$, with larger samples sizes leading to larger $\Delta\chi^2$.

Also similar to the tests of weak invariance results, a Strong Noninvariance Effect Size x Location of Noninvariance interaction accounted for variability in all of the $\Delta AFIs$, however the η^2 effect sizes were larger and ranged form 4.52-17.68%. Furthermore, a Strong Noninvariance Effect Size x Design Type interaction accounted for a notable amount of variability in each of the $\Delta AFIs$ with larger $\Delta AFIs$ being observed when the amount of noninvariance increased, and this effect was greater in the 3 (group) x 3 (time) design. Lastly, the η^2 effect sizes reported for the ΔCFI and ΔTLI were again very similar to the ΔCFI_A and ΔTLI_A, respectively.

Table 3.3: Percent Variance Explained by Study Conditions on the Change of Fit Indices for Tests of Strong Invariance

Predictors	$\Delta\chi^2$	$\Delta RMSEA$	ΔCFI	ΔCFI_A	ΔTLI	ΔTLI_A	$\Delta SRMR$
Sample Size (*N*)	**11.13**	0.81	0.01	0.01			0.12
Sample Size Ratio (R)	0.04	0.13	0.04	0.04	0.05	0.04	0.06
Strong Noninvariance Effect Size (SN)	**35.27**	**49.45**	**43.00**	**43.00**	**44.89**	**43.19**	**60.18**
Location of Noninvariance (LN)	**15.73**	**18.46**	**20.40**	**20.69**	**20.70**	**20.28**	**6.26**
Design Type (DT)	0.64	**4.74**	**3.64**	**3.63**	**3.05**	**4.17**	**3.38**
N x R							0.02
N x SN	**9.34**	0.17	0.02	0.02			1.24
R x SN	0.03	0.08	0.04	0.04	0.04	0.04	0.11
N x LN	**4.04**	0.01					0.10
R x LN	0.08	0.16	0.09	0.09	0.10	0.09	1.70
SN x LN	**13.04**	**8.31**	**17.68**	**17.65**	**17.13**	**16.50**	**4.52**
N x DT		0.08	0.01	0.01			0.07
R x DT							0.03
SN x DT		2.21	**3.09**	**3.01**	2.42	**3.26**	2.87
LN x DT	0.01	1.11	1.87	1.86	1.50	1.97	0.16
N x R x SN	0.01						0.01
N x R x LN	0.02						0.03
N x SN x LN	**3.35**	0.02					0.05
R x SN x LN	0.07	0.09	0.09	0.08	0.09	0.08	1.15
N x R x DT							0.01
N x SN x DT		0.02	0.01	0.01			0.13
R x SN x DT							0.03
N x LN x DT							
R x LN x DT							0.08
SN x LN x DT	0.01	0.40	1.53	1.49	1.22	1.56	0.08
N x R x SN x LN	0.02						0.01
N x R x SN x DT							
N x R x LN x DT							
N x SN x LN x DT		0.01					
R x SN x LN x DT							0.05
N x R x SN x LN x DT							

Note. Percent variance explained estimates are η^2 effect sizes. Model effects > 3% are in bold. Model effects that accounted for percent of variance explained < 0.01% are removed. Because the amount of change in *AIC* is not evaluated in model comparisons, ΔAIC is not included in the above table.

3.5 Power of $\Delta\chi^2$ and $\Delta AFIs$ for Tests of Invariance

The power of $\Delta AFIs$ for tests of weak and strong invariance across group or time was examined by calculating the percentage of converged replications in each study condition with noninvariance present that exceeded the condition specific cut-off values calculated in Phase 1 of the study (see above for details on cut-off value calculations). In Appendix B, Tables B.1 and B.2 contain the condition specific cut-off values for the 2 (group) x 2 (time) tests of weak and strong invariance, respectively, and Tables B.3 and B.4 contain cut-off values for the 3 (group) x 3 (time) tests of weak and strong invariance, respectively.

There were two exceptions to these power calculations, including calculations of power for the chi-square difference test and power for *AIC* tests. Because the $\Delta\chi^2$ is chi-square distributed, the critical value for a χ^2 distribution with $df = \Delta df = df_{constrained} - df_{unconstrained}$ and $\alpha = .05$ was used as the cut-off value and power was calculated as the percentage of $\Delta\chi^2$ that exceeded this critical value in conditions where noninvariance was present. Power for the *AIC* tests of invariance was calculated as the percentage of constrained model *AICs* that were lower than the unconstrained model *AICs* in each study condition.

3.5.1 $\Delta\chi^2$

Figure 3.1 displays the power of the chi-square difference test of weak invariance across group and time for each design type. Figure 3.2 displays the power of the chi-square difference test of strong invariance across group and time for each design type. Within each of the four conditions displayed in both Figures 3.1 and 3.2 larger sample sizes led to greater power to detect changes in both factor loadings and item intercepts. Likewise, as sample size ratio went from unbalanced (e.g., 4:1) to balanced (e.g, 1:1) the the power to detect a change in a factor loading or item intercept increased. Sample size and ratio showed greater influence on power for tests of weak and strong invariance across group.

Because a critical value from the χ^2 was used to calculate power, instead of deriving a cut-off

value from the observed distribution of $\Delta\chi^2$, the Type I error rate could also be examined. Type I error for the chi-square difference test was evaluated by examining the conditions where the $\Delta\lambda = 0$ or $\Delta\tau = 0$. Type I error for tests of weak invariance across time ranged from 4.4-8.2% for the 2 (group) x 2 (time) design and 5.00-5.01% for the 3 (group) x 3 (time) design, whereas Type I error for tests of weak invariance across group was 4.4-8.2% for the 2 (group) x 2 (time) design and 5.0-13.0% for the 3 (group) x 3 (time) design. Type I error rates for tests of strong invariance across time ranged from 4.2-7.8% for the 2 (group) x 2 (time) design and 5.00-5.01% for the 3 (group) x 3 (time) design, whereas Type I error rates for tests of strong invariance across group ranged from 4.2-7.8% in the 2 (group) x 2 (time) design and 4.8-6.8% in the 3 (group) x 3 (time) design.

3.5.2 $\Delta RMSEA$

Larger sample sizes led to greater power to detect changes in both factor loadings and items intercepts across groups or time. Surprisingly, in the 3 (group) x 3 (time) test of strong invariance condition, unbalanced samples showed slightly more power than balanced samples when the the total sample size (N) equaled 360, 720, and 1080. In the remaining conditions, power to detect a change in a factor loading or item intercept increased as sample size ratio went from unbalanced (e.g., 4:1) to balanced (e.g, 1:1). Results for $\Delta RMSEA$ are reported in Figures C.1 and C.2 in Appendix C.

3.5.3 ΔCFI and ΔCFI_A

Recall, in Table 3.1 the correlation between ΔCFI and ΔCFI_A was $r = .999$. This near-perfect correlation is demonstrated in these power calculations. Power was nearly identical for both tests of weak and strong invariance across study conditions for the ΔCFI and ΔCFI_A. The results from ΔCFI_A are discussed below. Results for ΔCFI are located in Appendix C.

Figure 3.3 displays power of the ΔCFI_A test of weak invariance across group and time for both design types. Figure 3.4 displays power of the ΔCFI_A test of strong invariance across group and

Figure 3.1: Power for $\Delta\chi^2$ Tests of Weak Invariance

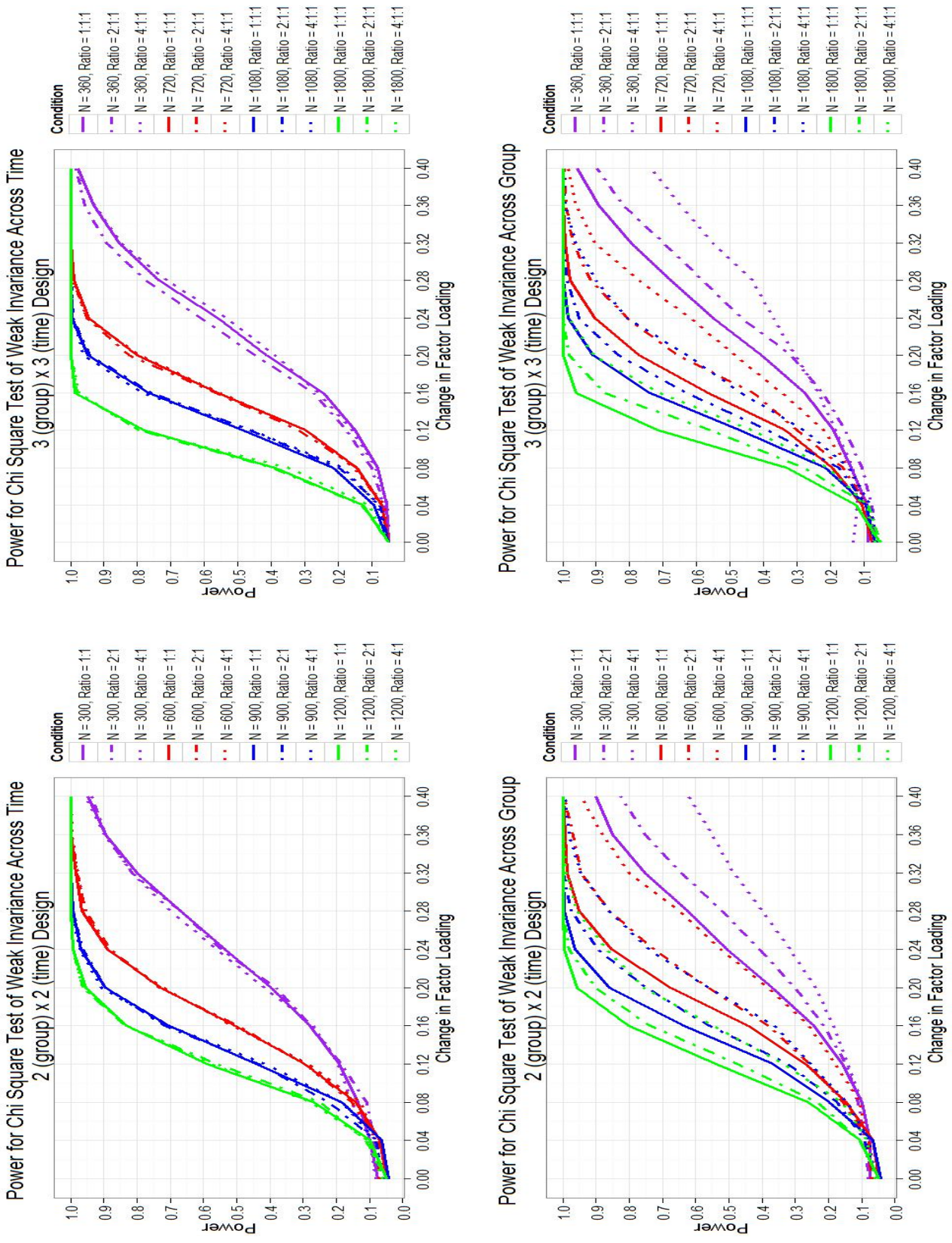
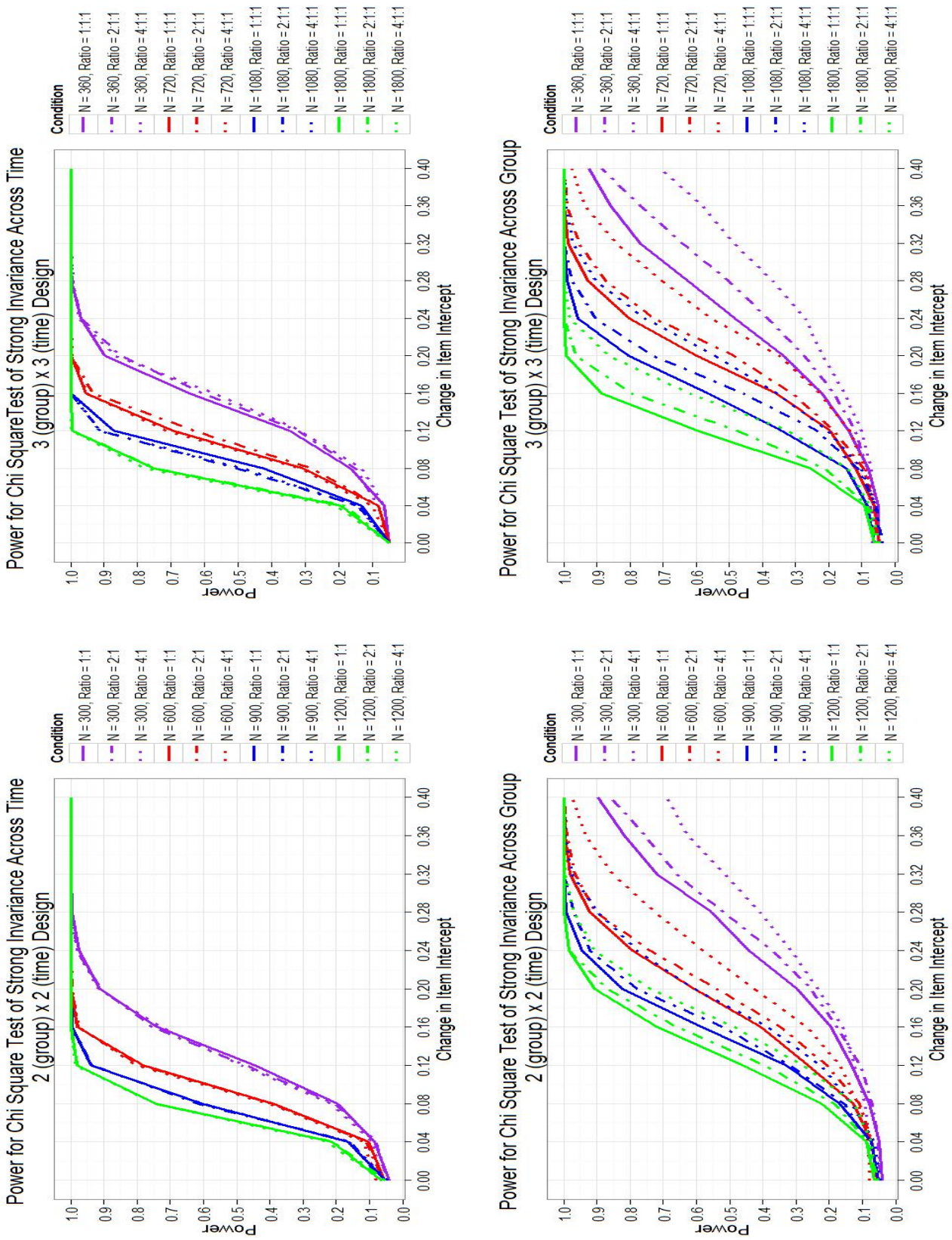


Figure 3.2: Power for $\Delta\chi^2$ Tests of Strong Invariance



time for both design types. Within each of the four conditions displayed in both Figures 3.3 and 3.4 larger sample sizes led to greater power to detect changes in both factor loadings and item intercepts. Likewise, as sample size ratio went from unbalanced (e.g., 4:1) to balanced (e.g, 1:1) the the power to detect a change in a factor loading or item intercept increased. In addition, for tests of weak and strong invariance across both group and time the 3 (group) x 3 (time) design demonstrated slightly higher power to detect a given effect size than the 2 (group) x 2 (time) design.

For instance, observe the condition where the n = 600 for each group (i.e., N = 1200, Ratio = 1:1, and N = 1800, Ratio = 1:1:1). Power to detect a $\Delta\lambda = 0.12$ for tests of weak invariance across time was .562 for the 2 (group) x 2 (time) design and .754 for the 3 (group) x 3 (time) design, whereas power for tests of weak invariance across group was .536 for the 2 (group) x 2 (time) design and .648 for the 3 (group) x 3 (time) design. This observed difference was not as pronounced for tests of strong invariance across time, where in the same sample size and ratio condition, power to detect a $\Delta\tau = .08$ was .668 for the 2 (group) x 2 (time) design and .660 for the 3 (group) x 3 (time) design, whereas power for tests of strong invariance across group was .176 for the 2 (group) x 2 (time) design and .226 for the 3 (group) x 3 (time) design.

3.5.4 ΔTLI and ΔTLI_A

Recall, in Table 3.1 the correlation between ΔTLI and ΔTLI_A was $r > .99$ for both tests of weak and strong invariance. This practically perfect correlation was apparent in these power calculations. Power was nearly identical for both tests of weak and strong invariance across study conditions for the ΔTLI and ΔTLI_A. Furthermore, ΔTLI and ΔTLI_A were highly correlated with ΔCFI and ΔCFI_A with correlations ranging from $r = .946 - .974$. Again power of ΔTLI and ΔTLI_A for tests of invariance was very similar to power of ΔCFI and ΔCFI_A tests. Results for ΔTLI and ΔTLI_A are included in Appendix C.

Figure 3.3: Power for ΔCFI_A Tests of Weak Invariance

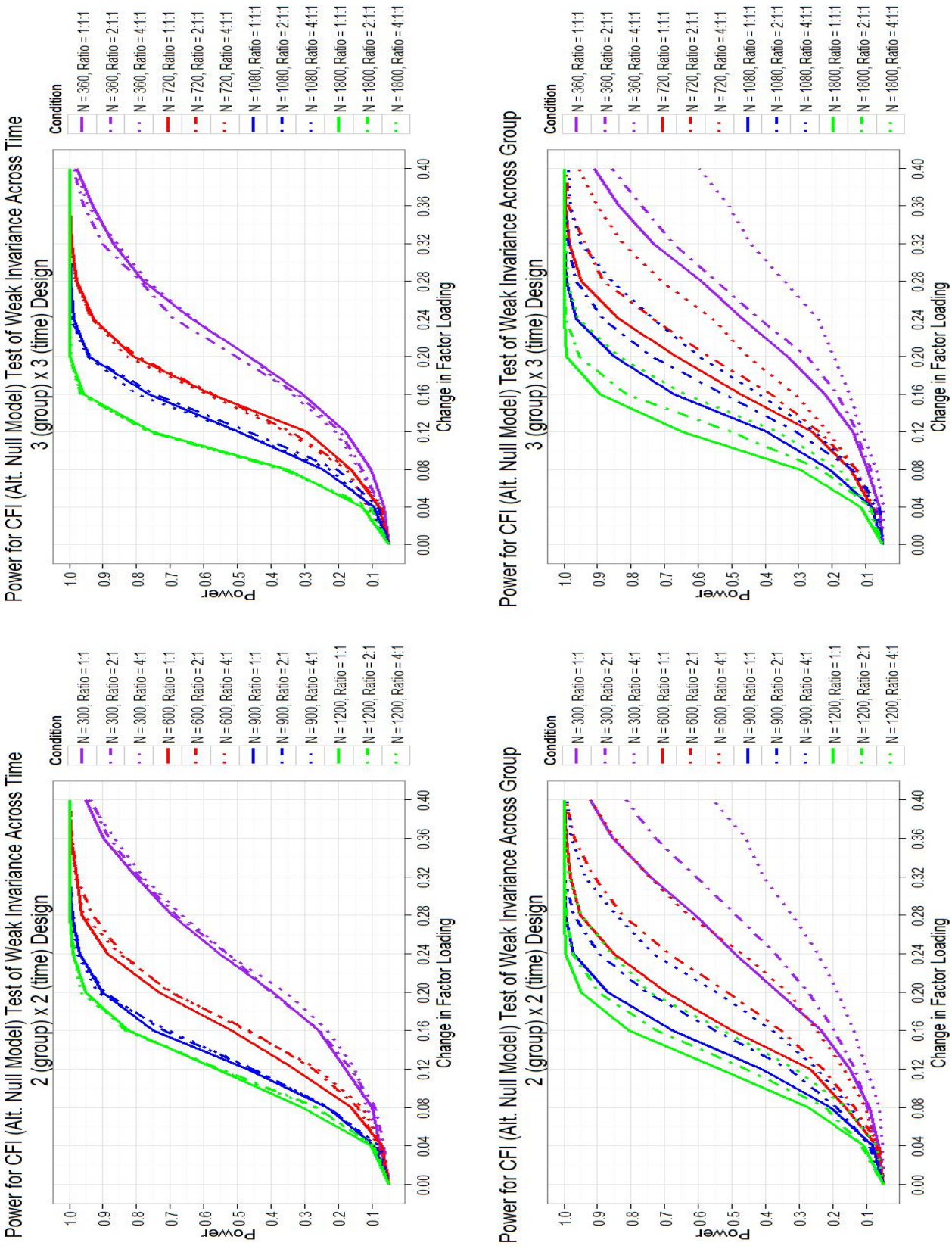
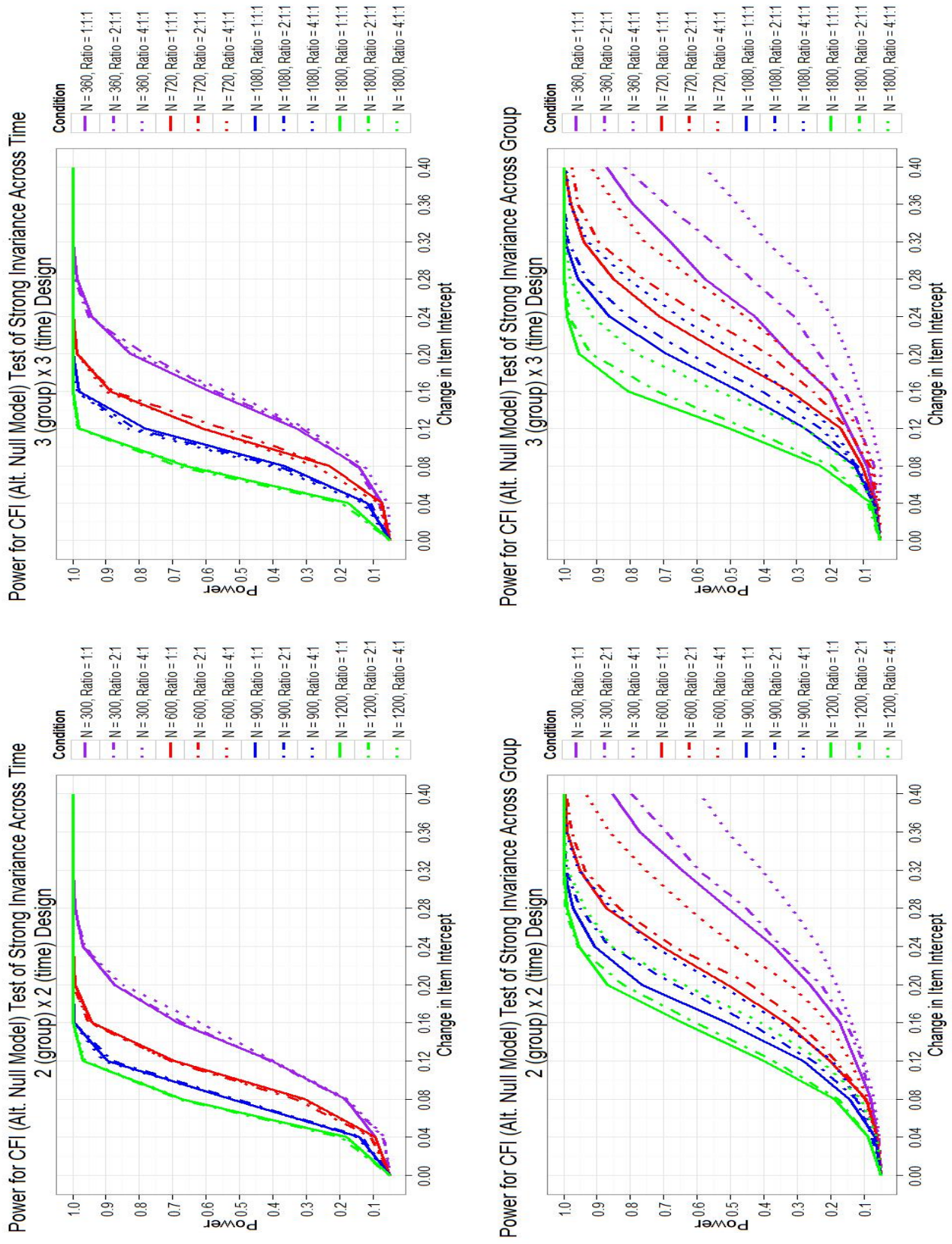


Figure 3.4: Power for ΔCFI_A Tests of Strong Invariance



3.5.5 $\Delta SRMR$

Figure 3.5 displays power of the $\Delta SRMR$ test of weak invariance across group and time for both design types. Figure 3.6 displays power of the $\Delta SRMR$ test of strong invariance across group and time for both design types. Similar to the previous $\Delta AFIs$, within each of the four conditions displayed in both Figures 3.5 and 3.6 larger sample sizes led to greater power to detect changes in both factor loadings and items intercepts.

As sample size ratio became more balanced the the power to detect a change in a factor loading increased for tests of invariance across group and also increased for tests of strong invariance across groups and time. Conversely, power decreased as sample size ratio became more balanced for tests of weak invariance across time.

3.5.6 AIC

Results for the power of AIC for tests of weak and strong invariance were similar in pattern to those observed with ΔCFI_A with some exceptions. Interestingly, within both tests of weak and strong invariance across both group and time the 2 (group) x 2 (time) design demonstrated slightly higher power to detect a given effect size than the 3 (group) x 3 (time) design. In fact, within 2 (group) x 2 (time) tests of weak invariance, the AIC had between 0 - .09 greater power than the ΔCFI_A, but in 3 (group) x 3 (time) design the ΔCFI_A demonstrated between 0 - .3 higher power, with the ΔCFI_A showing particularly higher power in conditions with smaller total sample sizes. Likewise, for tests of strong invariance, AIC demonstrated between 0- .124 higher power than ΔCFI_A in the 2 (group) x 2 (time) design, but between 0 - .26 lower power than ΔCFI_A in the 3 (group) x 3 (time) deign. Results for AIC are reported in Figures C.9 and C.10 in the Appendix C.

Figure 3.5: Power for $\Delta SRMR$ Tests of Weak Invariance

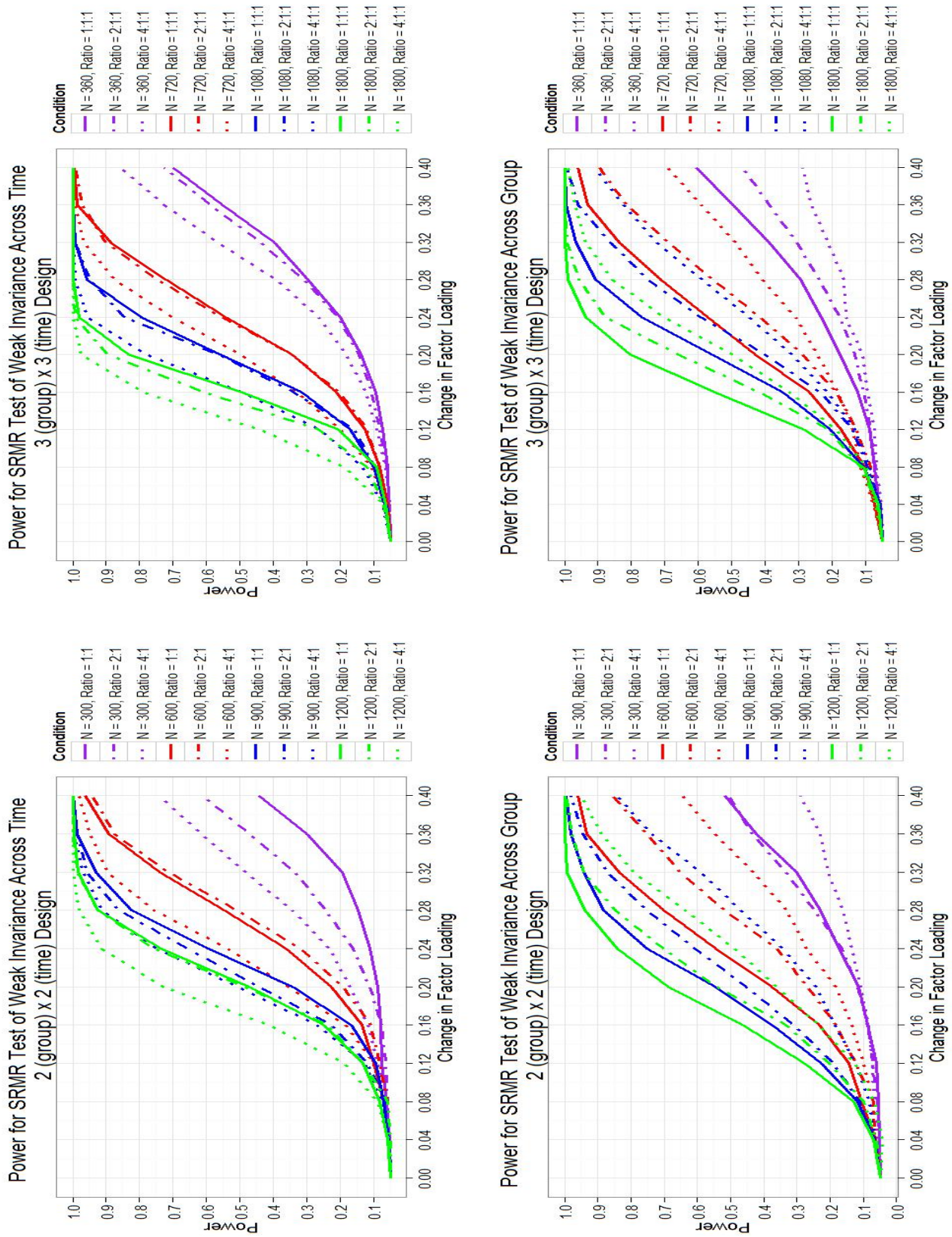
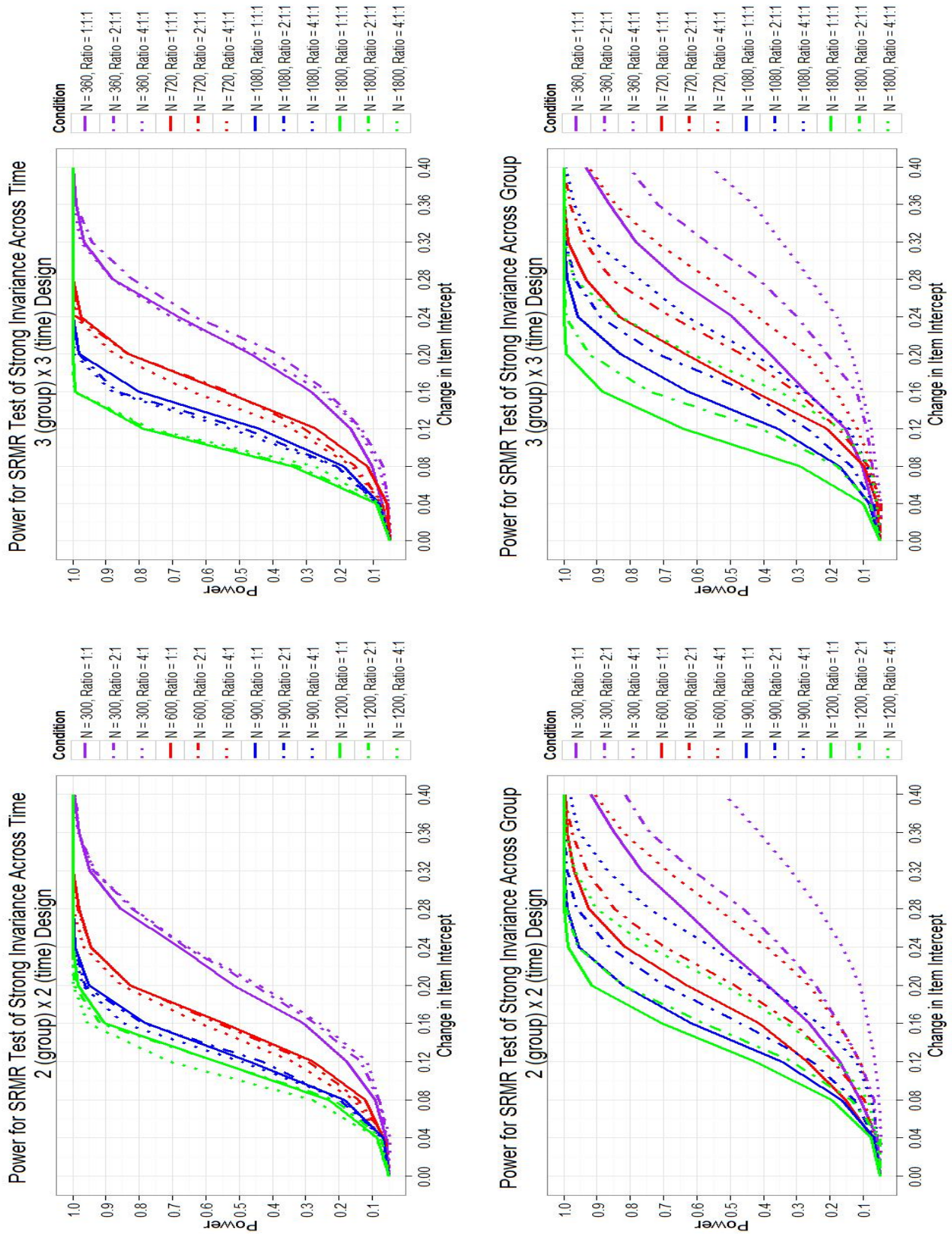


Figure 3.6: Power for Δ*SRMR* Tests of Strong Invariance



Chapter 4

Discussion

A common research question in psychology and the larger social sciences is if differences exist across groups or time on a measured construct. Prior to testing these hypotheses of group or longitudinal effects, researchers may first want to examine the psychometric properties of the measured construct in order to determine if the measure may be biased between groups or across time. Testing measurement invariance with multiple group and/or longitudinal CFA has a long history of use in psychology (Alwin & Jackson, 1981; French & Finch, 2006; Little, 1997, 2013; Marsh, 1994; Meredith, 1964, 1993; Meredith & Horn, 2001; Millsap, 2011; Reise et al., 1993; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000; Widaman et al., 2010). Measurement invariance in CFA is commonly evaluated by examining the change in model fit statistics as series of parameter constraints are made across groups or time.

Recently, researchers have been recommended to avoid using the $\Delta\chi^2$ for tests of measurement invariance because the test is overly sensitive to sample size (Brannick, 1995; Kelloway, 1995; Meade & Bauer, 2007). Instead, researchers have encouraged the examination of change in $AFIs$ for invariance tests (Cheung & Rensvold, 1999, 2002; Meade et al., 2008; Chen et al., 2008). Previous research on the use of $\Delta AFIs$ for tests of measurement invariance has examined the impact of a variety of study conditions (e.g., sample size, number of items per construct, factor loadings, factor correlation), but has been limited to scenarios of two groups measured at a single time point.

The present study expanded on this research by examining tests of weak and strong measurement invariance across multiple groups and time points. The current results are compared to previous research findings and discussed below.

4.1 Support for Hypotheses

4.1.1 Hypothesis 1

The first hypothesis stated power to detect weak or strong measurement bias would be lower for more complex (i.e., more groups and time points) mixed designs. Results from $\Delta\chi^2$and each of the $\Delta AFIs$ in the current study did not support hypothesis 1. Instead, power for tests weak and strong measurement invariance using $\Delta\chi^2$and each of the $\Delta AFIs$ was similar or slightly higher in the 3 (group) x 3 (time) design versus the 2 (group) x 2 (time) design. Conversely, use of the *AIC* for tests of invariance did support the hypothesis. Power of *AIC* for tests of invariance in 3 (group) x 3 (time) designs was as much as .25 lower when compared to tests of invariance in the 2 (group) x 2 (time) design. The calculation of the *AIC* can penalize model complexity and this may have been observed in the current study.

On the surface, the results for $\Delta AFIs$ may be potentially counter to previous research findings. Power for tests of measurement invariance has been reported to decrease as models became more complex, such as having more items (Chen, 2007; Meade et al., 2008) or factors (Meade et al., 2008). However, the current results may suggest that additional groups or time points are not the same kind of increase in model complexity that additional factors and items are.

A few possible explanations may exist for the current findings. First, although the study attempted to make total sample sizes, and group sample sizes similar across each design, the 3 (group) x 3 (time) design still had a larger total sample size than the 2 (group) x 2 (time) design. In addition, when noninvariance was present, it occurred more frequently (i.e., at both time 2 and time 3), possibly creating a larger overall effect size of measurement noninvariance. Both a larger sample size and/or effect size in the 3 (group) x 3 (time) design may have led to an increase in

power for tests of measurement invariance.

4.1.2 Hypothesis 2

The second hypothesis stated power to detect weak or strong measurement bias would decrease as the sample size ratio increases. Results from the current study partially supported this hypothesis. Power for tests of weak and strong invariance using $\Delta\chi^2$, ΔCFI, and ΔCFI_A was largest when sample sizes were balanced, and was inversely related to sample size ratio. These results supported Chen's (2007) reports that larger change in fit indices was observed for tests of invariance when sample sizes were balanced between groups than unbalanced. Likewise, in all conditions except the 3 (group) x 3 (time) tests of weak invariance, power for tests of weak and strong invariance using ΔTLI and ΔTLI_A was largest when sample sizes were balanced and decreased as sample size ratio increased. Excluding 2 (group) x 2 (time) and 3 (group) x 3 (time) tests of weak invariance across time, and the 3 (group) x 3 (time) test of strong invariance across time, power for tests of invariance using $\Delta RMSEA$ also was inversely related to sample size ratio. Although these few above conditions with tests using ΔTLI, ΔTLI_A, and $\Delta RMSEA$ did not support the sample size ratio hypothesis, the differences in power between the balanced and unbalanced designs was < .1 and may not be of much practical concern.

Results from the tests of weak and strong invariance using $\Delta SRMR$ were also mixed. Power for tests of strong invariance using $\Delta SRMR$ was largest when sample sizes were balanced, and decreased as sample size ratio increased. The same result was found for tests of weak invariance across group. Conversely, the opposite was found for tests of weak invariance across time. Unlike the above results, this unexpected difference in power was larger. In conditions where $\Delta\lambda > .32$, the unbalanced 4:1 and 4:1:1 designs demonstrated nearly .3 greater power than the balanced 1:1 and 1:1:1 sample size ratios. Overall, the $\Delta SRMR$ was the least related to the other fit indices, which is consistent with past research (Hu & Bentler, 1998; Meade et al., 2008) .

4.2 Additional Findings

Beyond the examined hypotheses, other notable results were revealed in the current study.

4.2.1 Total sample size versus sample size ratio

Data collection is frequently costly and time consuming, especially when multiple groups are being examined longitudinally. Researchers must carefully consider how large of a sample they can reasonably obtain. Furthermore, in many areas of social science, such as cross-cultural research, the focal group may be much smaller in size than the reference group in the population. In this all too common scenario the researcher may wonder if it is better to collect more individuals overall, or attempt to collect balanced samples. The current results provide some insight.

For example, examine the power for the tests of weak invariance across groups in the 2 (group) x 2 (time) design using ΔCFI, ΔTLI, ΔCFI_A, and ΔTLI_A. Power was very similar in the conditions where total sample size equaled 600, 900, and 1200 with sample size ratios of 1:1, 2:1, and 4:1, respectively. This pattern was also observed in other conditions and tests of invariance across groups. These results imply that if cost is a concern, it may be advantageous to collect a smaller total sample size that is balanced, than a larger total sample size where groups are unbalanced. If it is difficult to have the focal group sample size balanced with the reference group, researchers should be aware that a larger total sample size will be required to have increased power for tests of weak or strong invariance between groups.

4.2.2 Use of alternative null model

Current SEM software defaults to what is referred to as the independence null model for calculations of incremental fit indices including the CFI and TLI. Widaman and Thompson (2003) have suggested that this model is an inappropriate null model because it is not the worst fitting model to the data. Instead, a more restrictive alternative null model that includes equating variances of latent variables and item intercepts across each time point and group has been suggested for use in

calculating the CFI and TLI. The current study compared the use of the software default null and the alternative null model for tests of weak and strong invariance. Results revealed that the CFI and TLI were highly correlated with CFI_A and TLI_A, respectively. Power for tests of invariance using either null model were very similar. These results support an earlier example demonstration provided by Widaman and Thompson (2003) which also indicated ΔCFI and ΔTLI were very similar to ΔCFI_A and ΔTLI_A.

4.2.3 Current study cut-off values compared to previous recommendations

In the current study cut-off values for $\Delta AFIs$ were empirically derived by examining the distributions of $\Delta AFIs$ where invariance was present. Any variability in the $\Delta AFIs$ for these conditions was considered sampling variability and cut-off values at $\alpha = .05$ were determined by calculating either the 5th or 95th percentile for the fit index. As noted in the introduction, previous researchers (Cheung & Rensvold, 2002; Chen, 2007; Meade et al., 2008) have recommended several cut-off values for tests of weak or strong measurement invariance. Power for tests of invariance across the current study conditions was examined using these recommended cut-off values and results are presented in Appendix D.

The examination of previously recommended cut-off values yielded a few noteworthy results. First, the $\Delta CFI < .005$ cut-off for tests of weak invariance and $\Delta CFI < .002$ for tests of strong invariance recommended by Meade et al. (2008) led to increased Type I error rates. In particular, in the smallest sample size conditions (N = 300 or 360) the Type I error rate for tests of weak invariance exceeded .10 and exceeded .20 for tests of strong invariance. Next, Chen's (2007) recommended $\Delta SRMR$ values led to very low power across study conditions. Researchers are recommended to avoid using this $\Delta SRMR$ suggestion for tests of invariance. Lastly, Cheung and Rensvold's (2002) recommended $\Delta CFI < .01$ for tests of weak and strong invariance across time showed adequate power, but performed poorly for tests of invariance between groups in the current study's mixed designs. The cut-off values calculated for the current study (see Appendix B) may be more appropriate when researchers are testing invariance across both groups and time.

4.3 Limitations and Future Research

Several limitations and avenues for future research exist in the current study. First, the study contained only conditions where measurement bias was present between groups or across time. These conditions could be considered a main effect of group or time for the mixed design. The possible Group x Time interaction of measurement bias was not an examined condition. For example, suppose a researcher is conducting a pretest/post-test study for an experimental group and a control group (i.e., a 2 (group) x 2 (time) study). Further suppose that after the pretest the experimental manipulation creates measurement bias that is present in the post-test. Now, a Group x Time interaction would be present where the amount of measurement bias across time depends on group. Power to detect this noninvariance would likely be lower than power to detect the same level of noninvariance present in the post-test for both groups.

Certainly, this scenario may be of concern to researchers because tests of invariance may not reveal this Group x Time measurement bias, and the researcher may continue pursuing comparisons across group or time under the false assumption that measurement invariance is present. Future research should examine the power of ΔAFI tests to detect these possible interactions, and examine the possible consequences of assuming invariance and conducting additional tests (e.g., tests of factor variances/covariances, factor means, structural paths) when the interaction is present. Chen (2008) notes if weak or strong measurement noninvariance is present between the groups, but treated as invariant, then tests of latent regression paths and factor means for invariance across groups can lead to false group differences. The effect may be overestimated in one group and underestimated in another. Additional research is needed to better understand the consequences treating noninvariant data as invariant when testing additional parameters in mixed designs.

Another limitation in the current study is only uniform noninvariance was examined. Uniform noninvariance occurs when factor loadings or item intercepts only increase or decrease in one group or across time. For instance, in the current models, item 3 in the focal group only had decreases in factor loadings or increases in item intercepts. Mixed measurement bias would exist when some factor loadings or item intercepts in the focal group decreased, but other factor

loadings or item intercepts increased. Previous research (Chen, 2007; Meade & Bauer, 2007) on measurement invariance has examined the impact of uniform versus mixed measurement bias in tests of measurement invariance for the two group scenario, but future research should explore the effects of uniform versus mixed measurement bias in multiple group longitudinal designs.

Furthermore, when a test of measurement invariance fails at the weak or strong level, little guidance is provided for quantifying the effect size of this measurement bias (Millsap, 2005). Recently, Millsap and Olivera-Aguilar (2012) provided possible calculations of effect sizes for measurement bias in a two-group single time point CFA, but how large an effect must be to be considered practically significant instead of statistically significant remains unknown. Previous simulation studies have varied in the specified values of weak and strong measurement bias with factor loading differences between two groups being as great as 0.2 (Kim et al., 2012b; Millsap & Kwok, 2004), 0.25 (French & Finch, 2006, 2008; Meade & Lautenschlager, 2004), or 0.4 (Chen, 2007), and item intercept difference a great as 0.265 (Millsap & Kwok, 2004), or 0.4 (Chen, 2007). Interestingly, some of these selected values were simply the author's stated opinion of common differences found in social science research (e.g., Chen, 2007).

Future research should first begin by examining applications of measurement invariance testing in a specific research domain to determine what parameter differences may commonly exist in tests of weak and strong invariance. Undoubtedly, the definition for a "practically significant" amount of measurement bias will vary across psychological domains, with perhaps more conservative amounts existing in areas of high stakes testing. Thus, a single set of cut-off criteria for small, moderate, and large effect sizes will likely not be applicable across research domains. Instead, possible effect sizes for measurement bias should be influenced by a combination of substantive theory, previous research findings, and simulation results.

4.4 Conclusions

In conclusion, a few recommendations to researchers are offered. First, given the mixed sample size ratio results with the *SRMR*, the typically lower power to detect measurement invariance, and previous concerns about high variability in the fit index (Chen, 2007; Meade et al., 2008), researchers are encouraged to avoid using this fit statistic for evaluating tests of measurement invariance. Second, the *RMSEA* presented concerns similar to the *SRMR* and researchers may also want to avoid using this fit index for tests of invariance. Instead, the CFI and TLI showed to most promising results and researchers are encouraged to use these incremental fit indices. Third, use of the alternative null does not appear to impact tests of measurement invariance and researchers are encouraged to follow Widaman and Thompson's (2003) recommendations of using the alternative null model when investigating longitudinal data. Fourth, careful consideration of the trade-offs between total sample size and sample size ratio is needed, especially for tests of invariance across groups. The researcher may find it more efficient to collect more balanced samples than simply more data. Finally, the relationship between study conditions and tests for measurement invariance is complex. The current results shed more light on the impact of study choices on tests of measurement invariance using multiple group longitudinal CFA and highlighted areas for future research.

References

Aitken, A. (1935). Note on selection from a multivariate normal population. *Proceedings of the Edinburgh Mathematical Society (Series 2)*, 4, 106–110.

Akaike, H. (1987). Factor analysis and aic. *Psychometrika*, 52, 317–322.

Alwin, D. F. & Jackson, D. J. (1981). Applications of simultaneous factor analysis to issues of factorial invariance. In D. Jackson & E. Borgatta (Eds.), *Factor analysis and measurement in sociological research: A multi-dimensional perspective* (pp. 249–279). Beverly Hills, CA: Sage.

Anderson, J. C. & Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika*, 49, 155–173.

Atienza, F. L., Balaguer, I., & Garcia-Merita, M. L. (2003). Satisfaction with life scale: Analysis of factorial invariance across sexes. *Personality and Individual Differences*, 35, 1255–1260.

Bagozzi, R. P. (1977). Structural equation models in experimental research. *Journal of Marketing Research*, 14, 209–226.

Bagozzi, R. P., Yi, Y., & Singh, S. (1991). On the use of structural equation models in experimental designs: Two extensions. *International Journal of Research in Marketing*, 8, 125–140.

Bandalos, D. L. (2006). The role of simulation in structural equation modeling. In G. R. Hancock & R. Mueller (Eds.), *Structural Equaltion Mdoeling: A Second Course* (pp. 385–426). Grenwich, CT: Information Age.

Bandalos, D. L. & Gange, P. (2012). Simulation methods in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 92–108). New York: Guilford Press.

Barbosa-Leiker, C., Kostick, M., Lei, M., McPherson, S., Roper, V., Hoekstra, T., & Wright, B. (2013). Measurement invariance of the perceived stress scale and latent mean differences across gender and time. *Stress and Health*, 29, 253–260.

Bauer, D. J. (2005). The role of nonlinear factor-to-indicator relationships in tests of measurement equivalence. *Psychological Methods*, 10, 305–316.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107, 238–246.

Bowers, E. P., Li, Y., Kiely, M. K., Brittian, A., Lerner, J. V., & Lerner, R. M. (2010). The five c's model of positive youth development: A longitudinal analysis of confirmatory factor structure and measurement invariance. *Journal of Youth and Adolescence*, 39, 720–735.

Brannick, M. T. (1995). Critical comments on applying covariance structure modeling. *Journal of Organizational Behavior*, 16, 201–213.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford Press.

Burnham, K. P. & Anderson, D. R. (2004). Multimodel inference understanding aic and bic in model selection. *Sociological Methods & Research*, 33, 261–304.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, 105, 456–466.

Chen, F., Bollen, K. A., Paxton, P., Curran, P. J., & Kirby, J. B. (2001). Improper solutions

in structural equation models causes, consequences, and strategies. *Sociological Methods & Research*, 29, 468–508.

Chen, F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in rmsea test statistic in structural equation models. *Sociological Methods & Research*, 36, 462–494.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, 14, 464–504.

Chen, F. F. (2008). What happens if we compare chopsticks with forks? the impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology*, 95, 1005–1018.

Chen, F. F., Sousa, K. H., & West, S. G. (2005). Teacher's corner: Testing measurement invariance of second-order factor models. *Structural Equation Modeling*, 12, 471–492.

Cheung, G. W. & Lau, R. S. (2012). A direct comparison approach for testing measurement invariance. *Organizational Research Methods*, 15, 167–198.

Cheung, G. W. & Rensvold, R. B. (1998). Cross-cultural comparisons using non-invariant measurement items. *Applied Behavioral Science Review*, 6, 93–110.

Cheung, G. W. & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management*, 25, 1–27.

Cheung, G. W. & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9, 233–255.

Coertjens, L., Donche, V., De Maeyer, S., Vanthournout, G., & Van Petegem, P. (2012). Longitudinal measurement invariance of likert-type learning strategy scales are we using the same ruler at each wave? *Journal of Psychoeducational Assessment*, 30, 577–587.

Cole, D. A. & Maxwell, S. E. (1985). Multitrait-multimethod comparisons across populations: A confirmatory factor analytic approach. *Multivariate Behavioral Research*, 20, 389–417.

Crawford, J. R. & Henry, J. D. (2004). The positive and negative affect schedule (panas): Construct validity, measurement properties and normative data in a large non-clinical sample. *British Journal of Clinical Psychology*, 43, 245–265.

De Ayala, R. J. (2009). *Theory and practice of item response theory*. New York: Guilford Press.

DeShon, R. P. (1998). A cautionary note on measurement error corrections in structural equation models. *Psychological Methods*, 3, 412–423.

DeShon, R. P. (2004). Measures are not invariant across groups without error variance homogeneity. *Psychology Science*, 46, 137–149.

Dimitrov, D. M. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counseling and Development*, 43, 121–149.

French, B. F. & Finch, W. H. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling*, 13, 378–402.

French, B. F. & Finch, W. H. (2008). Multigroup confirmatory factor analysis: Locating the invariant referent sets. *Structural Equation Modeling*, 15, 96–113.

Genz, A. & Bretz, F. (2009). *Computation of multivariate normal and t probabilities. Lecture notes in statistics*, volume 195. Heidelberg: Springer-Verlage.

Gerbing, D. W. & Anderson, J. C. (1987). Improper solutions in the analysis of covariance structures: Their interpretability and a comparison of alternate respecifications. *Psychometrika*, 52, 99–111.

Glanville, J. L. & Wildhagen, T. (2007). The measurement of school engagement assessing dimensionality and measurement invariance across race and ethnicity. *Educational and Psychological Measurement*, 67, 1019–1041.

Guttman, L. (1954). Some necessary conditions for common-factor analysis. *Psychometrika*, 19, 149–161.

Hancock, G. R. (1997). Structural equation modeling methods of hypothesis testing of latent variable means. *Measurement and Evaluation in Counseling and Development*, 20, 91–105.

Hancock, G. R. (2001). Effect size, power, and sample size determination for structured means modeling and mimic approaches to between-groups hypothesis testing of means on a single latent construct. *Psychometrika*, 66, 373–388.

Hancock, G. R. (2004). Experimental, quasi-experimental, and nonexperimental design and analysis with latent variables. *The Sage handbook of quantitative methodology for the social sciences*, (pp. 317–334).

Hancock, G. R., Lawrence, F. R., & Nevitt, J. (2000). Type 1 error and power of latent mean methods and manova in factorially invariant and noninvariant latent variable systems. *Structural Equation Modeling*, 7, 534–556.

Heine, S. J. & Hamamura, T. (2007). In search of east asian self-enhancement. *Personality and Social Psychology Review*, 11, 4–27.

Horn, J. L. & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, 18, 117–144.

Hu, L. T. & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, 3, 424–453.

Hu, L. T. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6(1), 1–55.

Jöreskog, K. G. (1970). Estimation and testing of simplex models. *British Journal of Mathematical and Statistical Psychology*, 23, 121–145.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, 36, 409–426.

Kaplan, D. (2009). *Structural equation modeling: Foundations and extensions*, volume 10. Thousand Oaks, CA: Sage.

Kelloway, E. K. (1995). Structural equation modelling in perspective. *Journal of Organizational Behavior*, 16, 215–224.

Kim, E. S., Kwok, O. M., & Yoon, M. (2012a). Testing factorial invariance in multilevel data: A monte carlo study. *Structural Equation Modeling*, 19, 250–267.

Kim, E. S., Yoon, M., & Lee, T. (2012b). Testing measurement invariance using mimic likelihood ratio test with a critical value adjustment. *Educational and Psychological Measurement*, 72, 469–492.

Kline, R. B. (2011). *Principles and practice of structural equation modeling*. New York: Guilford press.

Little, T. D. (1997). Mean and covariance structures (macs) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research*, 32, 53–76.

Little, T. D. (2013). *Longitudinal structural equation modeling*. New York: Guilford Press.

Little, T. D., Card, N. A., Slegers, D. W., & Ledford, E. C. (2007a). Representing contextual effects in multiple-group macs models. In T. D. Little, B. J. A., & N. A. Card (Eds.), *Modeling contextual effects in longitudinal studies* (pp. 121–147). Mahwah, NJ: LEA.

Little, T. D., Preacher, K. J., Selig, J. P., & Card, N. A. (2007b). New developments in latent variable panel analyses of longitudinal data. *International Journal of Behavioral Development*, 31, 357–365.

Marsh, H. W. (1994). Confirmatory factor analysis models of factorial invariance: A multifaceted approach. *Structural Equation Modeling*, 1, 5–34.

Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating cfa and efa: Application to students' evaluations of university teaching. *Structural Equation Modeling*, 16, 439–476.

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology*, 60, 577–605.

Meade, A. W. & Bauer, D. J. (2007). Power and precision in confirmatory factor analytic tests of measurement invariance. *Structural Equation Modeling*, 14, 611–635.

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, 93, 568–592.

Meade, A. W. & Lautenschlager, G. J. (2004). A monte-carlo study of confirmatory factor analytic tests of measurement equivalence/invariance. *Structural Equation Modeling*, 11, 60–72.

Meredith, W. (1964). Notes on factorial invariance. *Psychometrika*, 29, 177–185.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58, 525–543.

Meredith, W. & Horn, J. L. (2001). The role of factorial invariance in modeling growth and change. In A. Sayer & L. M. Collins (Eds.), *New Methods for the Analysis of Change* (pp. 203–240). Washington D. C: American Psychological Association.

Millsap, R. E. (2005). Four unresolved problems in studies of factorial invariance. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics* (pp. 153–172). Mahwah, NJ: Lawrence Erlbaum Associates.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York: Routledge.

Millsap, R. E. & Kwok, O. M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods*, 9, 93.

Millsap, R. E., Meredith, W., Cudeck, R., & MacCallum, R. (2007). Factorial invariance: Historical perspectives and new problems. In R. Cudeck & R. C. MacCallum (Eds.), *Factor Analysis at 100* (pp. 131–152). Mahwah, NJ: Lawrence Erlbaum Associates.

Millsap, R. E. & Olivera-Aguilar, M. (2012). Investigating measurement invariance using confirmatory factor analysis. In R. H. Hoyle (Ed.), *Handbook of Structural Equation Modeling* (pp. 380–392). New York: Guilford Press.

Muthén, L. K. & Muthén, B. O. (2010). *Mplus: Statistical analysis with latent variables: User's guide*. Los Angles: Muthén & Muthén.

Oort, F. J. (1998). Simulation study of item bias detection with restricted factor analysis. *Structural Equation Modeling*, 5, 107–124.

Pentz, M. A. & Chou, C.-P. (1994). Measurement invariance in longitudinal clinical research assuming change from development and intervention. *Journal of Consulting and Clinical Psychology*, 62, 450–462.

Pitts, S. C., West, S. G., & Tein, J.-Y. (1996). Longitudinal measurement models in evaluation research: Examining stability and change. *Evaluation and Program Planning*, 19, 333–350.

Ployhart, R. E. & Oswald, F. L. (2004). Applications of mean and covariance structure analysis: Integrating correlational and experimental approaches. *Organizational Research Methods*, 7, 27–65.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: a comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology*, 87, 517–529.

Raykov, T. (2005). Studying group and time invariance in maximal reliability for multiple-component measuring instruments via covariance structure modelling. *British Journal of Mathematical and Statistical Psychology*, 58, 301–317.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychological Bulletin*, 114, 552–566.

Rensvold, R. B. & Cheung, G. W. (2001). Testing for metric invariance using structural equation models: Solving the standardization problem. *Research in Management*, 1, 21–50.

Rosseel, Y. (2012). lavaan: An r package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1–36.

Sass, D. A. (2011). Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework. *Journal of Psychoeducational Assessment*, 29, 347–363.

Satorra, A. & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, 66, 507–514.

Savalei, V. & Kolenikov, S. (2008). Constrained versus unconstrained estimation in structural equation modeling. *Psychological Methods*, 13, 150–170.

Schmitt, N. & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review*, 18, 210–222.

Short, S. D. & Hawley, P. H. (2012). Evolutionary attitudes and literacy survey (eals): Development and validation of a short form. *Evolution: Education and Outreach*, 5, 419–428.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology*, 27, 229–239.

Steenkamp, J. B. E. & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, 25, 78–107.

Steiger, J. H. (1989). *EzPATH: causal modeling*. Evanston, IL: Systat.

Tanka, J. S. (1993). Multifaceted conceptions of fit in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 10–39). Newbury Park, CA: Sage.

Thurstone, L. L. (1947). *Multiple Factor Analysis*. Chicago: Chicago University Press.

Tucker, L. R. & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38, 1–10.

van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, 9, 486–492.

Vandenberg, R. J. & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3, 4–70.

Wicherts, J. M. & Dolan, C. V. (2004). A cautionary note on the use of information fit indexes in covariance structure modeling with means. *Structural Equation Modeling*, 11, 45–50.

Wicherts, J. M. & Dolan, C. V. (2010). Measurement invariance in confirmatory factor analysis: An illustration using iq test performance of minorities. *Educational Measurement: Issues and Practice*, 29, 39–47.

Wicherts, J. M., Dolan, C. V., Hessen, D. J., Oosterveld, P., Van Baal, G. C. M., Boomsma, D. I., & Span, M. M. (2004). Are intelligence tests measurement invariant over time? investigating the nature of the flynn effect. *Intelligence*, 32, 509–537.

Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child Development Perspectives*, 4, 10–18.

Widaman, K. F. & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281–324). Washington D. C: American Psychological Association.

Widaman, K. F. & Thompson, J. S. (2003). On specifying the null model for incremental fit indices in structural equation modeling. *Psychological methods*, 8, 16–37.

Woods, C. M. & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Applied Psychological Measurement*, 35, 339–361.

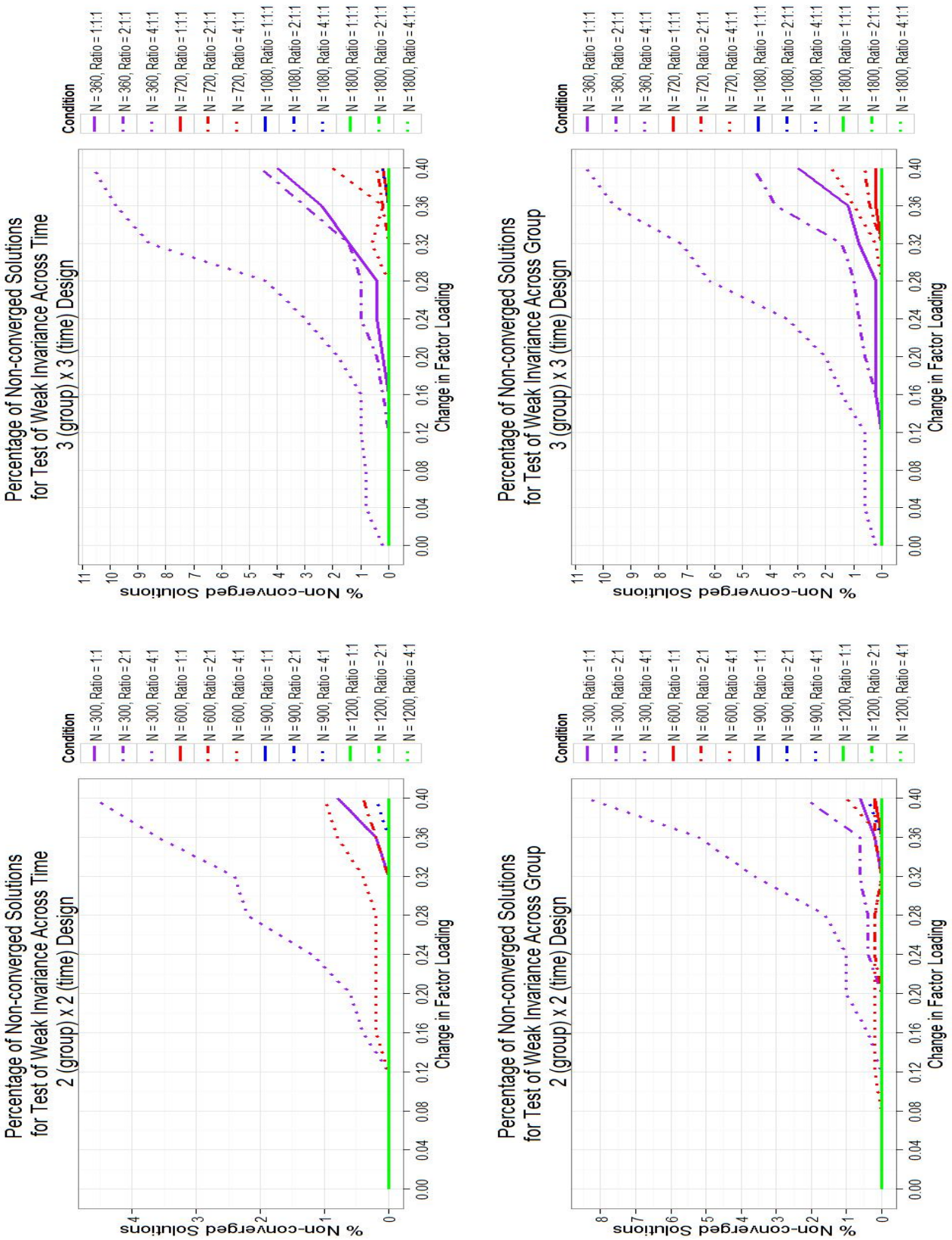
Wu, C.-H., Chen, L. H., & Tsai, Y.-M. (2009). Longitudinal invariance analysis of the satisfaction with life scale. *Personality and individual differences*, 46, 396–401.

Yoon, M. & Millsap, R. E. (2007). Detecting violations of factorial invariance using data-based specification searches: A monte carlo study. *Structural Equation Modeling*, 14, 435–463.

Appendix A

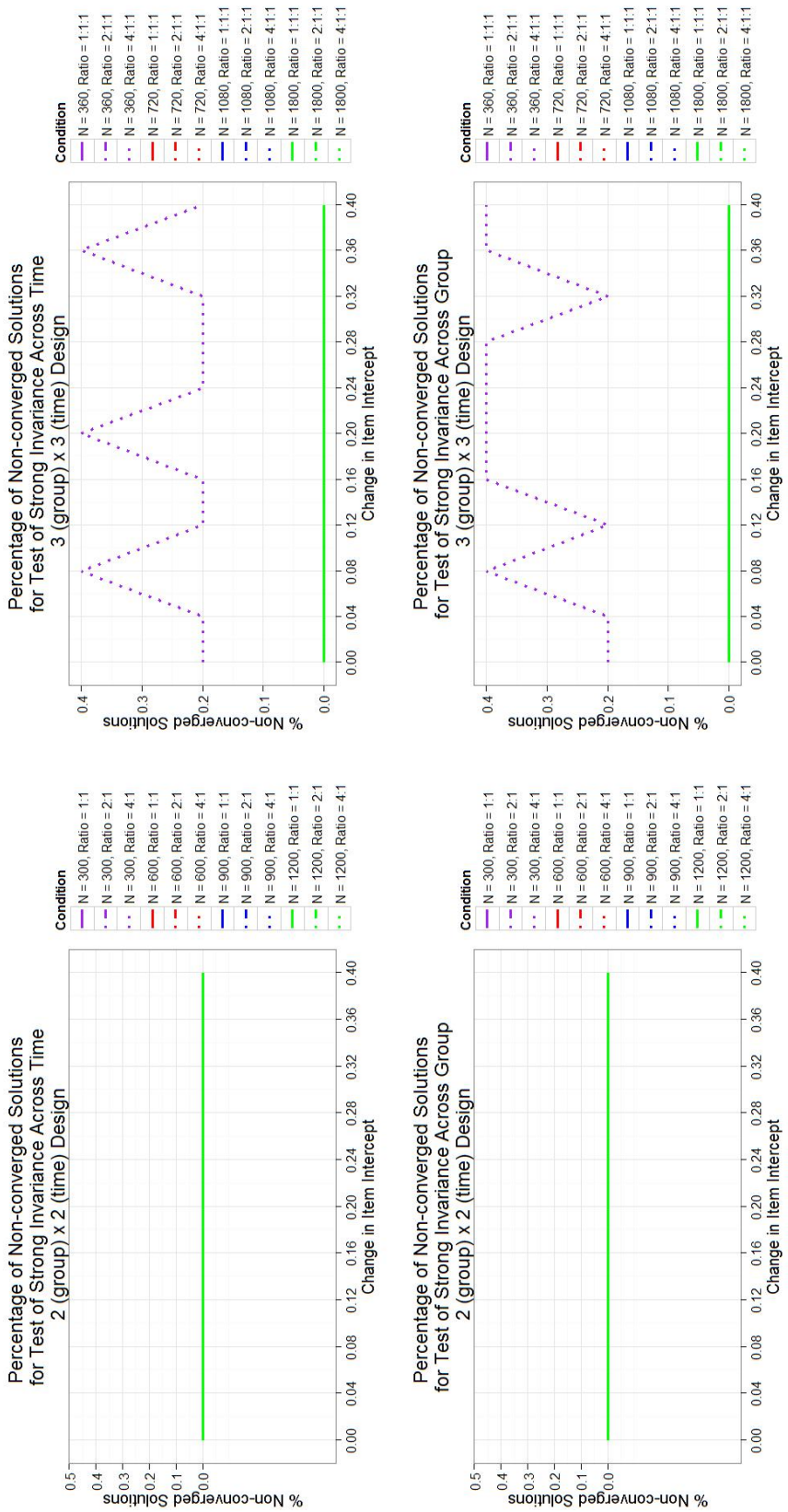
Percentage of Non-converged and Improper Solutions

Figure A.1: Percentage of Non-converged Solutions for Tests of Weak Invariance



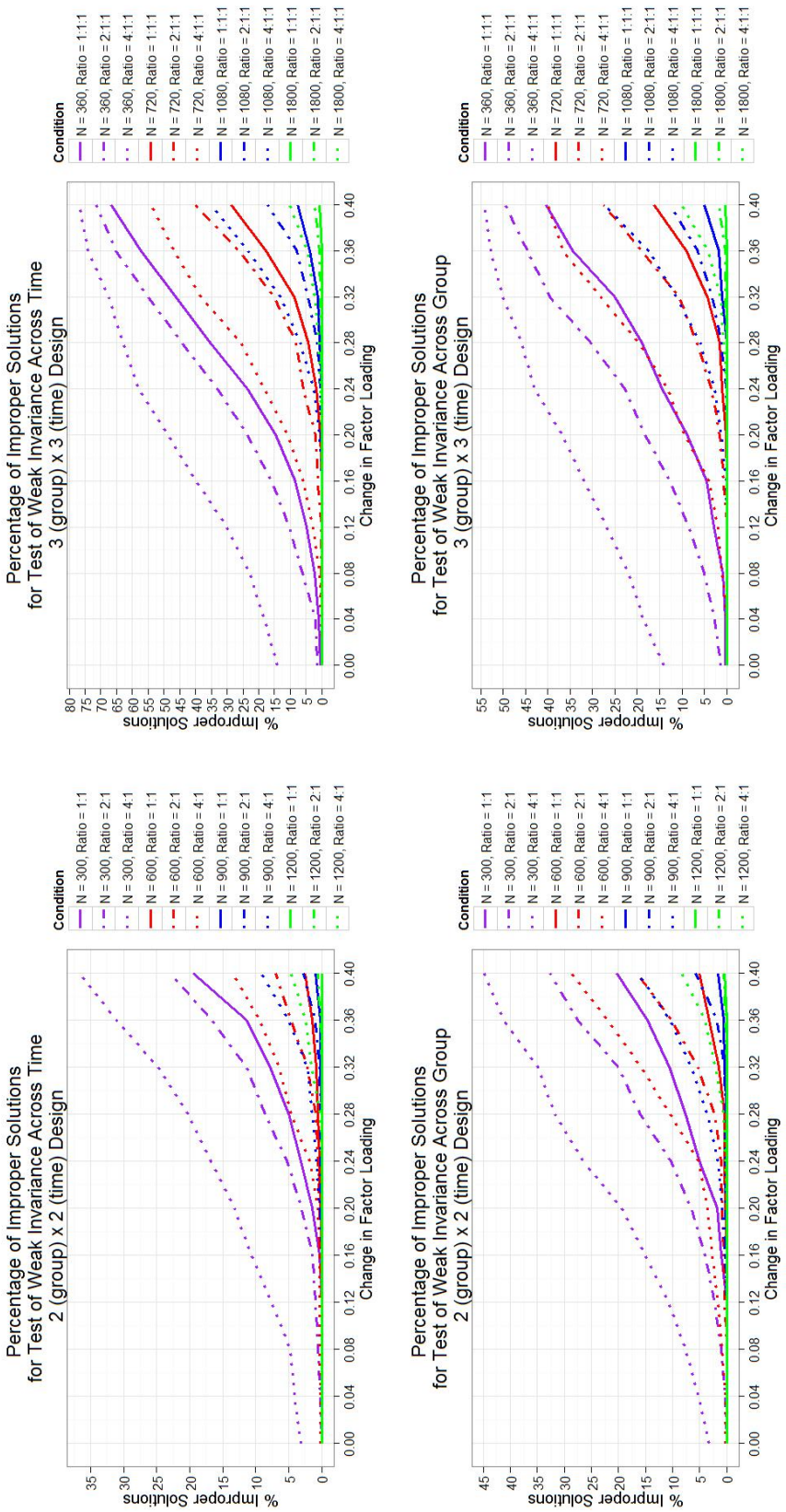
Note. Lines for conditions not present in the above plots overlap with the solid line at 0%

Figure A.2: Percentage of Non-converged Solutions for Tests of Strong Invariance



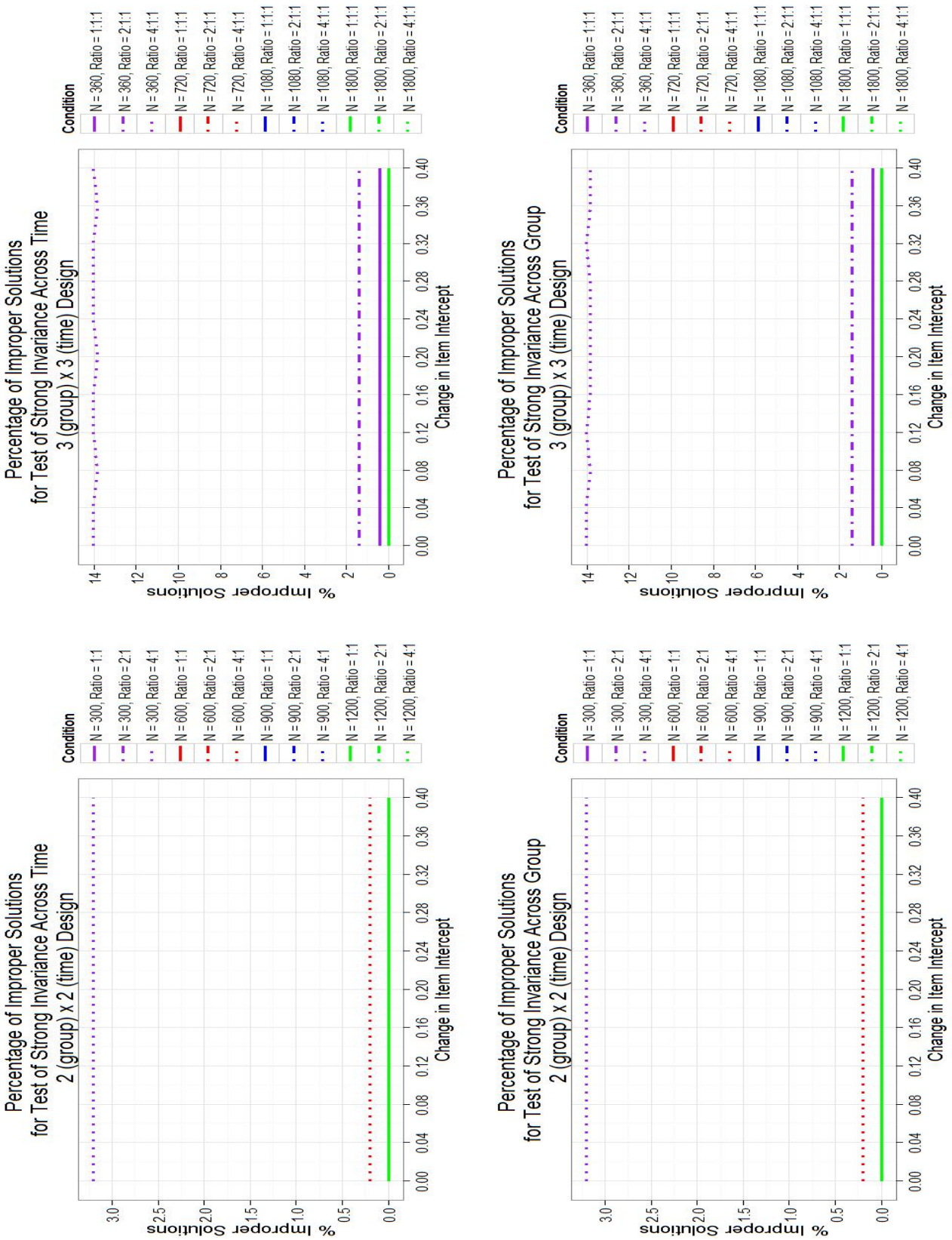
Note. Lines for conditions not present in the above plots overlap with the solid line at 0%

Figure A.3: Percentage of Improper Solutions for Tests of Weak Invariance



Note. Lines for conditions not present in the above plots overlap with the solid line at 0%

Figure A.4: Percentage of Improper Solutions for Tests of Strong Invariance



Note. Lines for conditions not present in the above plots overlap with the solid line at 0%.

Appendix B

Cut-off Values for $\Delta AFIs$

Table B.1: Cut-off Values for 2 (group) x 2 (time) Test of Weak Invariance

N	$\Delta RMSEA$	ΔCFI	ΔCFI_A	ΔTLI	ΔTLI_A	$\Delta SRMR$
Sample Size Ratio = 1:1						
300	0.040	-0.010	-0.009	-0.023	-0.037	0.041
600	0.026	-0.004	-0.004	-0.011	-0.017	0.025
900	0.019	-0.002	-0.002	-0.007	-0.011	0.020
1200	0.021	-0.002	-0.002	-0.006	-0.009	0.018
Sample Size Ratio = 2:1						
300	0.040	-0.010	-0.010	-0.024	-0.038	0.035
600	0.030	-0.005	-0.005	-0.013	-0.020	0.026
900	0.024	-0.003	-0.003	-0.008	-0.012	0.019
1200	0.019	-0.002	-0.002	-0.005	-0.009	0.018
Sample Size Ratio = 4:1						
300	0.040	-0.011	-0.011	-0.024	-0.038	0.031
600	0.028	-0.005	-0.005	-0.012	-0.019	0.021
900	0.022	-0.003	-0.003	-0.007	-0.012	0.018
1200	0.019	-0.002	-0.002	-0.006	-0.009	0.015

Note. N = total sample size. Cut-off values for $\Delta RMSEA$ and $\Delta SRMR$ are the 95^{th} percentiles of the sampling distribution when weak invariance was present. Cut-off values for ΔCFI, ΔCFI_A, ΔTLI, and ΔTLI_A are the 5^{th} percentiles of the sampling distribution when weak invariance was present.

Table B.2: Cut-off Values for 2 (group) x 2 (time) Test of Strong Invariance

N	$\Delta RMSEA$	ΔCFI	ΔCFI_A	ΔTLI	ΔTLI_A	$\Delta SRMR$
Sample Size Ratio = 1:1						
300	0.025	-0.009	-0.009	-0.013	-0.020	0.010
600	0.020	-0.005	-0.005	-0.007	-0.010	0.008
900	0.017	-0.003	-0.003	-0.005	-0.008	0.006
1200	0.016	-0.002	-0.002	-0.004	-0.006	0.006
Sample Size Ratio = 2:1						
300	0.028	-0.009	-0.009	-0.013	-0.021	0.012
600	0.018	-0.004	-0.004	-0.007	-0.011	0.008
900	0.018	-0.003	-0.003	-0.005	-0.008	0.007
1200	0.014	-0.002	-0.002	-0.004	-0.006	0.006
Sample Size Ratio = 4:1						
300	0.031	-0.010	-0.010	-0.015	-0.024	0.013
600	0.021	-0.005	-0.005	-0.009	-0.014	0.009
900	0.018	-0.003	-0.003	-0.005	-0.008	0.007
1200	0.015	-0.002	-0.002	-0.004	-0.006	0.006

Note. N = total sample size. Cut-off values for $\Delta RMSEA$ and $\Delta SRMR$ are the 95$^{\text{th}}$ percentiles of the sampling distribution when strong invariance was present. Cut-off values for ΔCFI, ΔCFI_A, ΔTLI, and ΔTLI_A are the 5$^{\text{th}}$ percentiles of the sampling distribution when strong invariance was present.

Table B.3: Cut-off Values for 3 (group) x 3 (time) Test of Weak Invariance

N	$\Delta RMSEA$	ΔCFI	ΔCFI_A	ΔTLI	ΔTLI_A	$\Delta SRMR$
Sample Size Ratio = 1:1:1						
360	0.029	-0.008	-0.007	-0.018	-0.025	0.024
720	0.019	-0.004	-0.004	-0.008	-0.011	0.015
1080	0.016	-0.002	-0.002	-0.005	-0.007	0.012
1800	0.010	-0.001	-0.001	-0.003	-0.004	0.010
Sample Size Ratio = 2:1:1						
360	0.026	-0.007	-0.007	-0.015	-0.022	0.022
720	0.021	-0.003	-0.003	-0.008	-0.012	0.015
1080	0.016	-0.002	-0.002	-0.005	-0.007	0.012
1800	0.011	-0.001	-0.001	-0.003	-0.004	0.009
Sample Size Ratio = 4:1:1						
360	0.034	-0.008	-0.008	-0.017	-0.025	0.021
720	0.018	-0.003	-0.003	-0.008	-0.012	0.013
1080	0.014	-0.002	-0.002	-0.005	-0.008	0.010
1800	0.011	-0.001	-0.001	-0.003	-0.004	0.008

Note. N = total sample size. Cut-off values for $\Delta RMSEA$ and $\Delta SRMR$ are the 95^{th} percentiles of the sampling distribution when weak invariance was present. Cut-off values for ΔCFI, ΔCFI_A, ΔTLI, and ΔTLI_A are the 5^{th} percentiles of the sampling distribution when weak invariance was present.

Table B.4: Cut-off Values for 3 (group) x 3 (time) Test of Strong Invariance

N	$\Delta RMSEA$	ΔCFI	ΔCFI_A	ΔTLI	ΔTLI_A	$\Delta SRMR$
Sample Size Ratio = 1:1:1						
360	0.021	-0.006	-0.006	-0.010	-0.015	0.009
720	0.014	-0.003	-0.003	-0.005	-0.007	0.006
1080	0.012	-0.002	-0.002	-0.004	-0.005	0.005
1800	0.009	-0.001	-0.001	-0.002	-0.003	0.004
Sample Size Ratio = 2:1:1						
360	0.017	-0.007	-0.007	-0.011	-0.015	0.010
720	0.011	-0.003	-0.003	-0.005	-0.008	0.006
1080	0.009	-0.002	-0.002	-0.003	-0.004	0.005
1800	0.009	-0.001	-0.001	-0.002	-0.003	0.004
Sample Size Ratio = 4:1:1						
360	0.018	-0.008	-0.008	-0.010	-0.015	0.010
720	0.014	-0.003	-0.003	-0.005	-0.008	0.007
1080	0.012	-0.002	-0.002	-0.003	-0.005	0.005
1800	0.009	-0.001	-0.001	-0.002	-0.003	0.005

Note. N = total sample size. Cut-off values for $\Delta RMSEA$ and $\Delta SRMR$ are the 95$^{\text{th}}$ percentiles of the sampling distribution when strong invariance was present. Cut-off values for ΔCFI, ΔCFI_A, ΔTLI, and ΔTLI_A are the 5$^{\text{th}}$ percentiles of the sampling distribution when strong invariance was present.

Appendix C

Additional Results for Power of $\Delta AFIs$ for Tests of Invariance

C.1 Results for $\Delta RMSEA$

Figure C.1 displays power of the $\Delta RMSEA$ test of weak invariance across group and time for both design types. Figure C.2 displays power of the $\Delta RMSEA$ test of strong invariance across group and time for both design types.

C.2 Results for ΔCFI

Figure C.3 displays power of the ΔCFI test of weak invariance across group and time for both design types. Figure C.4 displays power of the ΔCFI test of strong invariance across group and time for both design types. In each of the four conditions displayed in Figures C.3 and C.4 larger sample sizes led to greater power to detect changes in both factor loadings and items intercepts. Similarly, as sample size ratio went from unbalanced (e.g., 4:1) to balanced (e.g, 1:1) the the power to detect a change in a factor loading or item intercept increased. In addition, for tests of weak and strong invariance across both group and time the 3 (group) x 3 (time) design demonstrated slightly higher power to detect a given effect size than the 2 (group) x 2 (time) design.

Figure C.1: Power for $\Delta RMSEA$ Tests of Weak Invariance

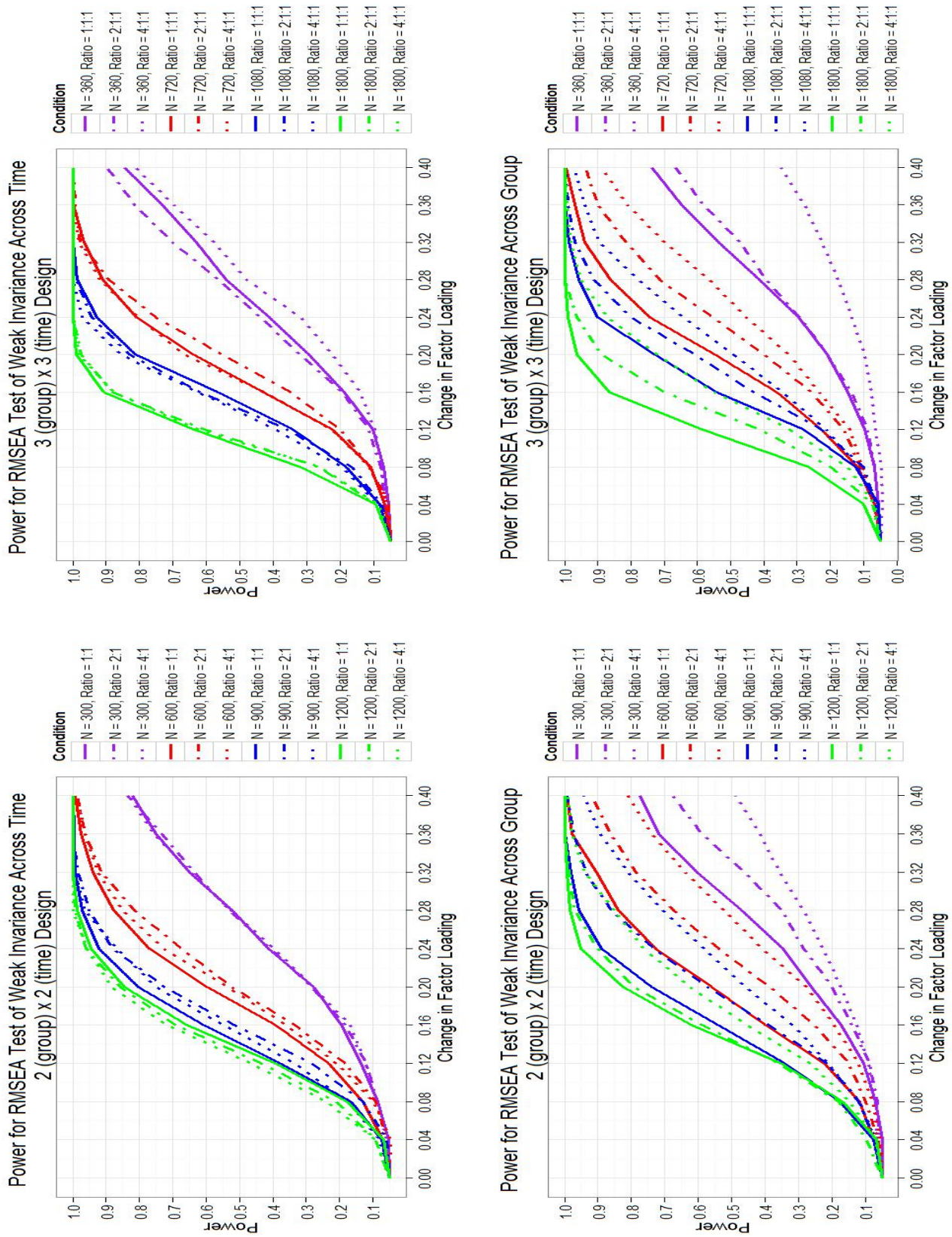
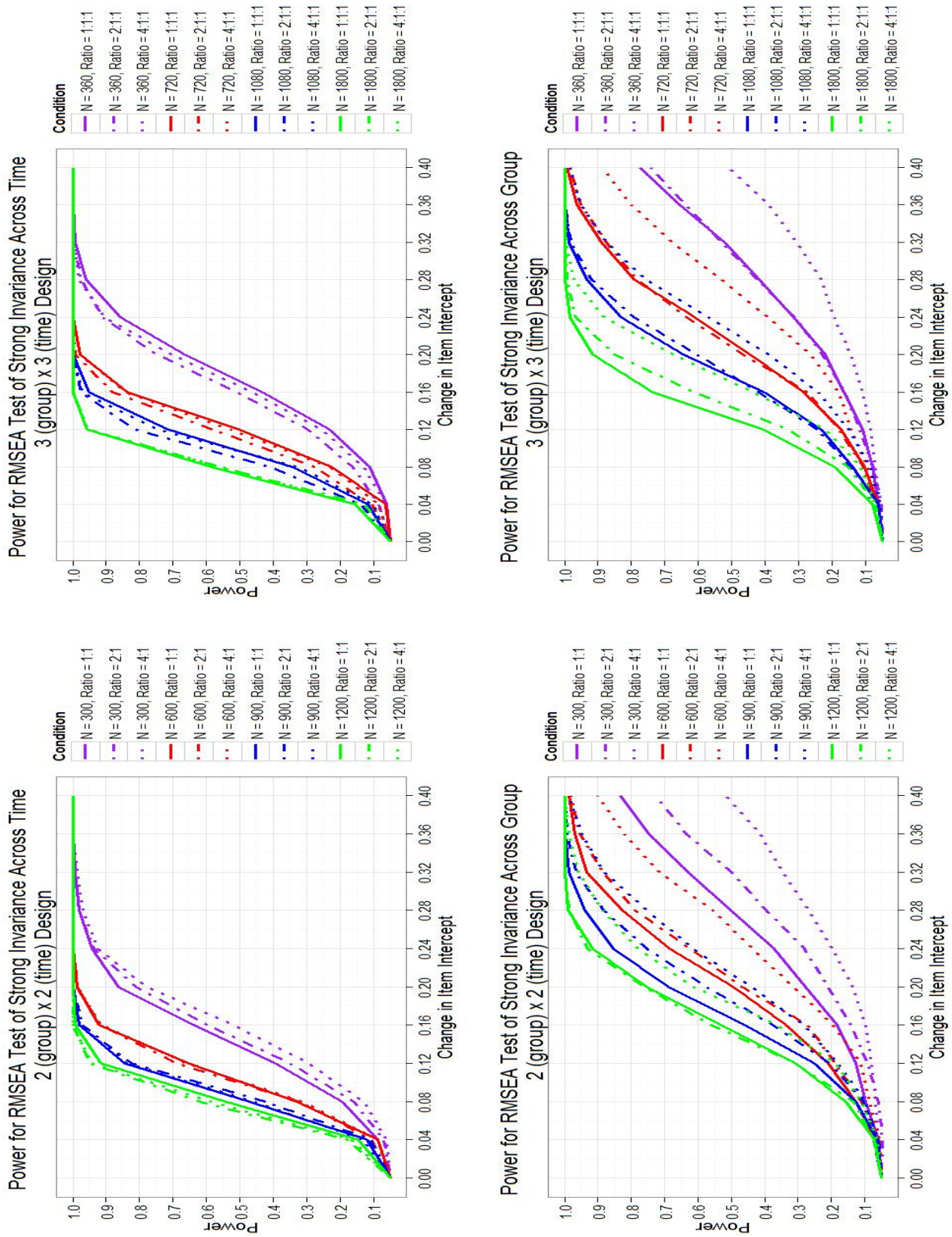


Figure C.2: Power for $\Delta RMSEA$ Tests of Strong Invariance



For example, examine the conditions where the n = 600 for each group (i.e., N = 1200, Ratio = 1:1, and N = 1800, Ratio = 1:1:1). Power to detect a $\Delta\lambda = 0.12$ for tests of weak invariance across time was .562 for the 2 (group) x 2 (time) design and .752 for the 3 (group) x 3 (time) design, whereas power for tests of weak invariance across group was .536 for the 2 (group) x 2 (time) design and .646 for the 3 (group) x 3 (time) design. This observed difference was not as dramatic for tests of strong invariance, where in the same sample size and ratio condition, power to detect a $\Delta\tau = .08$ was .668 for the 2 (group) x 2 (time) design and .660 for the 3 (group) x 3 (time) design, whereas power for tests of strong invariance across group was .176 for the 2 (group) x 2 (time) design and .224 for the 3 (group) x 3 (time) design.

C.3 Results for ΔTLI

Figure C.5 displays power of the ΔTLI test of weak invariance across group and time for both design types. Figure C.6 displays power of the ΔTLI test of strong invariance across group and time for both design types. In each of the four conditions displayed in Figures C.5 and C.6 larger sample sizes led to greater power to detect changes in both factor loadings and items intercepts. Similarly, as sample size ratio went from unbalanced (e.g., 4:1) to balanced (e.g, 1:1) the the power to detect a change in a factor loading or item intercept increased. The impact of sample size ratio on power was more apparent in tests of invariance between groups than across time. In addition, for tests of weak and strong invariance across both group and time the 3 (group) x 3 (time) design demonstrated slightly higher power to detect a given effect size than the 2 (group) x 2 (time) design.

For example, examine the condition where the n = 600 for each group (i.e., N = 1200, Ratio = 1:1, and N = 1800, Ratio = 1:1:1). Power to detect a $\Delta\lambda = 0.12$ for tests of weak invariance across time was .534 for the 2 (group) x 2 (time) design and .772 for the 3 (group) x 3 (time) design, whereas power for tests of weak invariance across group was .488 for the 2 (group) x 2 (time) design and .682 for the 3 (group) x 3 (time) design. This observed difference was not as large for tests of strong invariance across time, where in the same sample size and ratio condition, power to

Figure C.3: Power for ΔCFI Tests of Weak Invariance

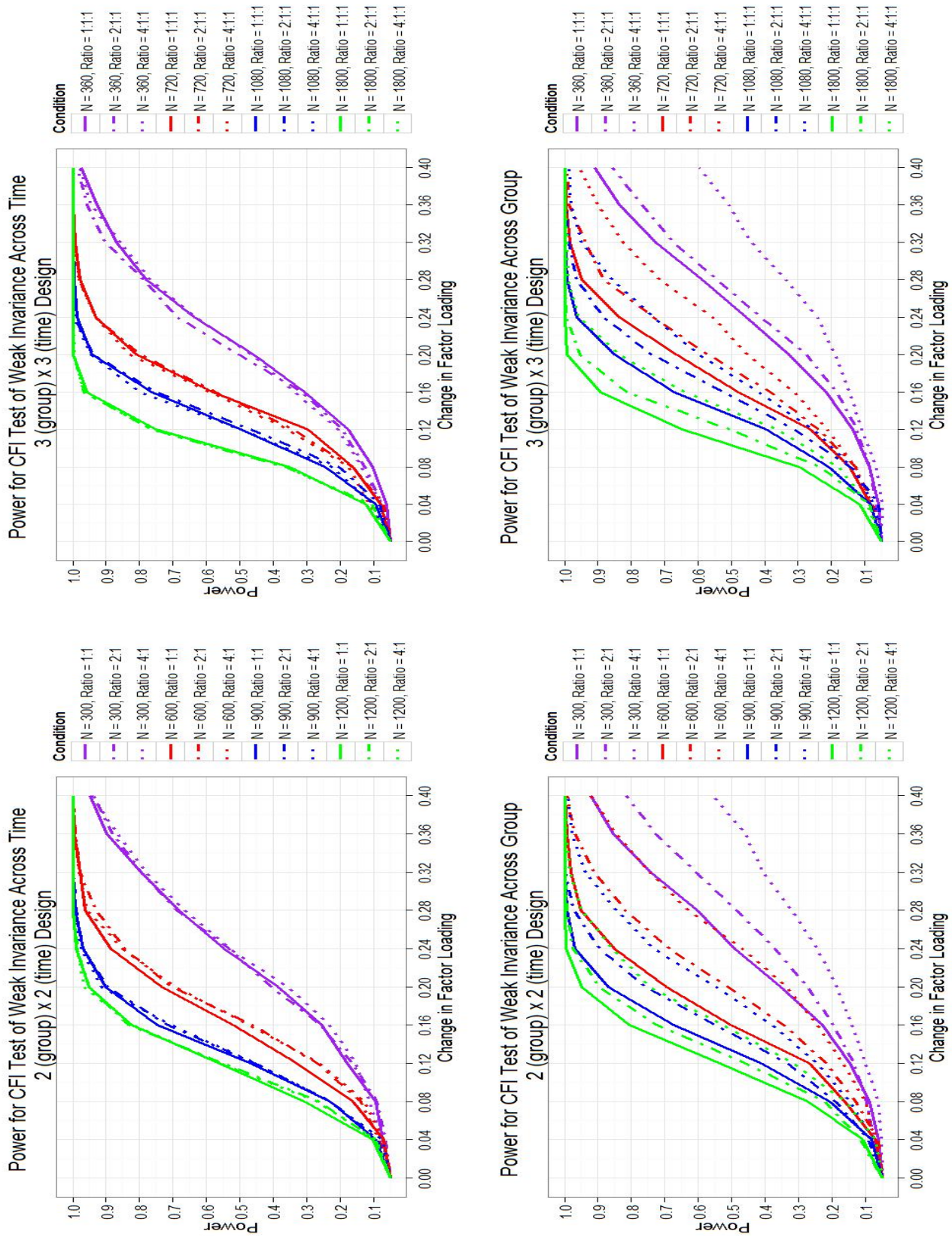
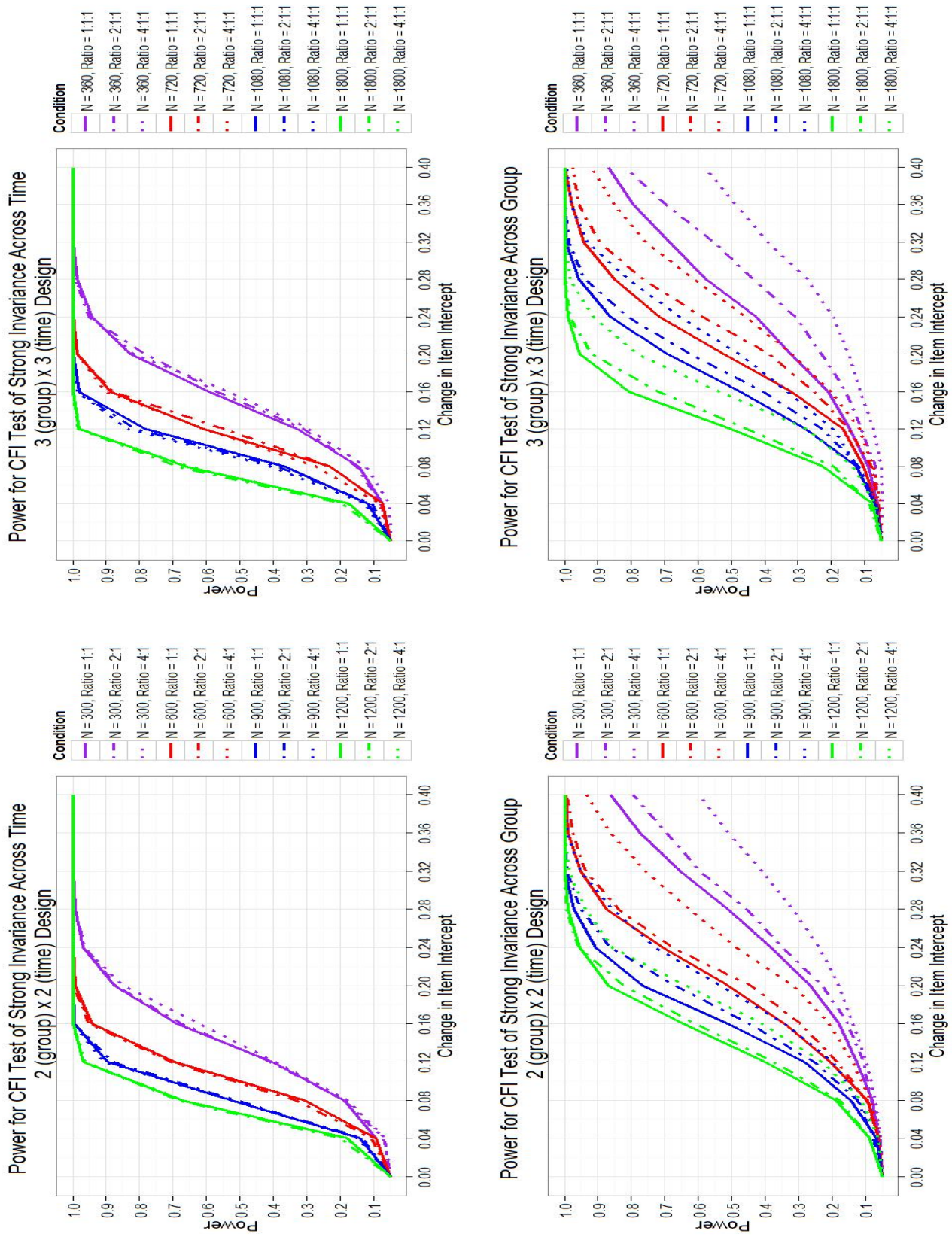


Figure C.4: Power for $\Delta\grave{C}FI$ Tests of Strong Invariance



detect a $\Delta\tau = .08$ was .658 for the 2 (group) x 2 (time) design and .676 for the 3 (group) x 3 (time) design, whereas power for tests of strong invariance across group was .188 for the 2 (group) x 2 (time) design and .218 for the 3 (group) x 3 (time) design.

C.4 Results for ΔTLI_A

Figure C.7 displays power of the ΔTLI_A test of weak invariance across group and time for both design types. Figure C.8 displays power of the ΔTLI_A test of strong invariance across group and time for both design types. Similar to the ΔTLI, within each of the four conditions displayed in both Figures C.7 and C.8 larger sample sizes led to greater power to detect changes in both factor loadings and items intercepts. Likewise, as sample size ratio went from unbalanced (e.g., 4:1) to balanced (e.g, 1:1) the the power to detect a change in a factor loading or item intercept increased. Much the same as the ΔTLI, the effect of sample size ratio on power was more noticeable in tests of invariance between groups than across time. Furthermore, within both tests of weak and strong invariance across both group and time the 3 (group) x 3 (time) design demonstrated slightly higher power to detect a given effect size than the 2 (group) x 2 (time) design.

For example, within the condition where the n = 600 for each group (i.e., N = 1200, Ratio = 1:1, and N = 1800, Ratio = 1:1:1). Power to detect a $\Delta\lambda = 0.12$ for tests of weak invariance across time was .534 for the 2 (group) x 2 (time) design and .772 for the 3 (group) x 3 (time) design, whereas power for tests of weak invariance across group was .486 for the 2 (group) x 2 (time) design and .682 for the 3 (group) x 3 (time) design. This observed difference was not as pronounced for tests of strong invariance across time, where in the same sample size and ratio condition, power to detect a $\Delta\tau = .08$ was .660 for the 2 (group) x 2 (time) design and .676 for the 3 (group) x 3 (time) design, whereas power for tests of strong invariance across group was .186 for the 2 (group) x 2 (time) design and .218 for the 3 (group) x 3 (time) design.

Figure C.5: Power for ΔTLI Tests of Weak Invariance

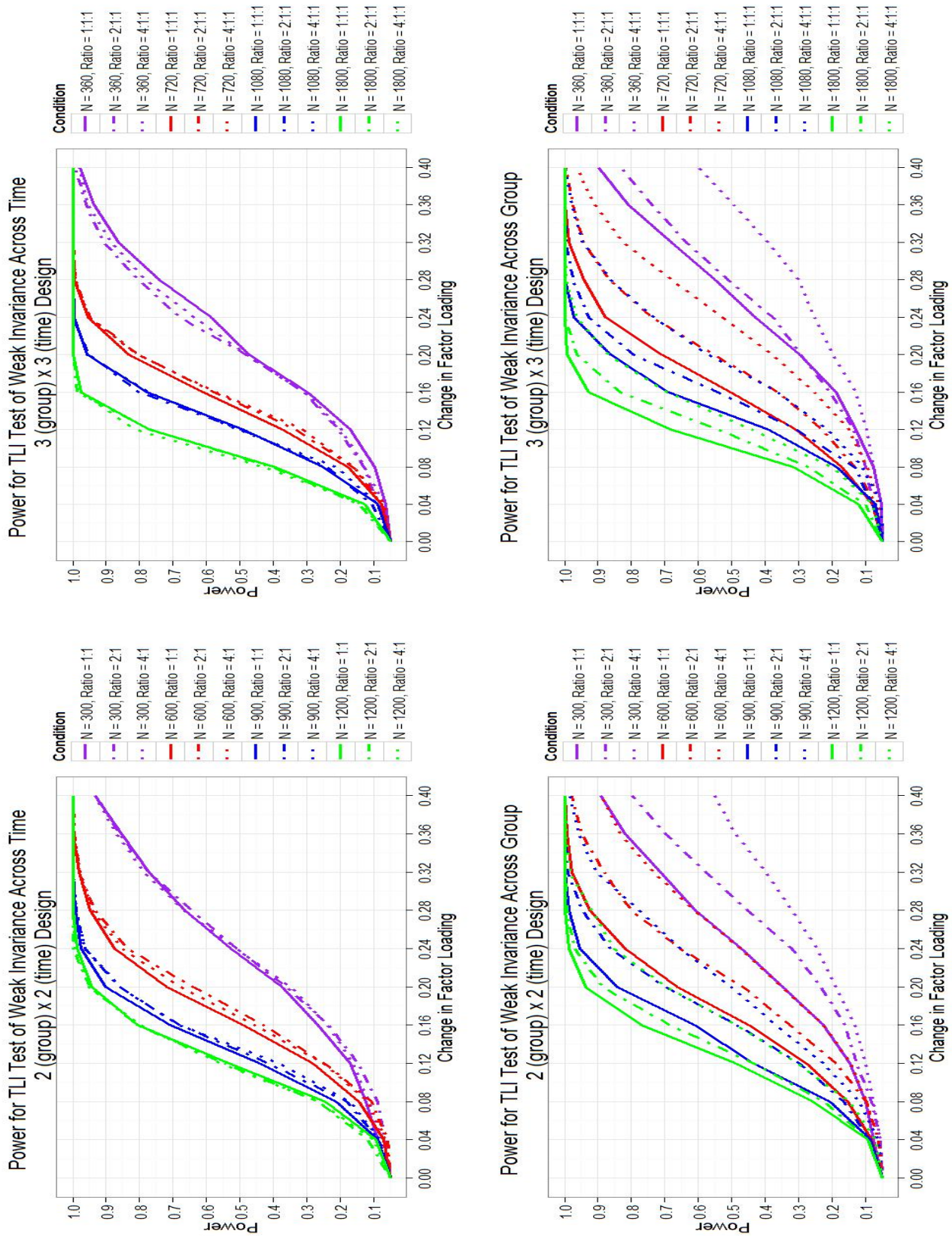


Figure C.6: Power for ΔTLI Tests of Strong Invariance

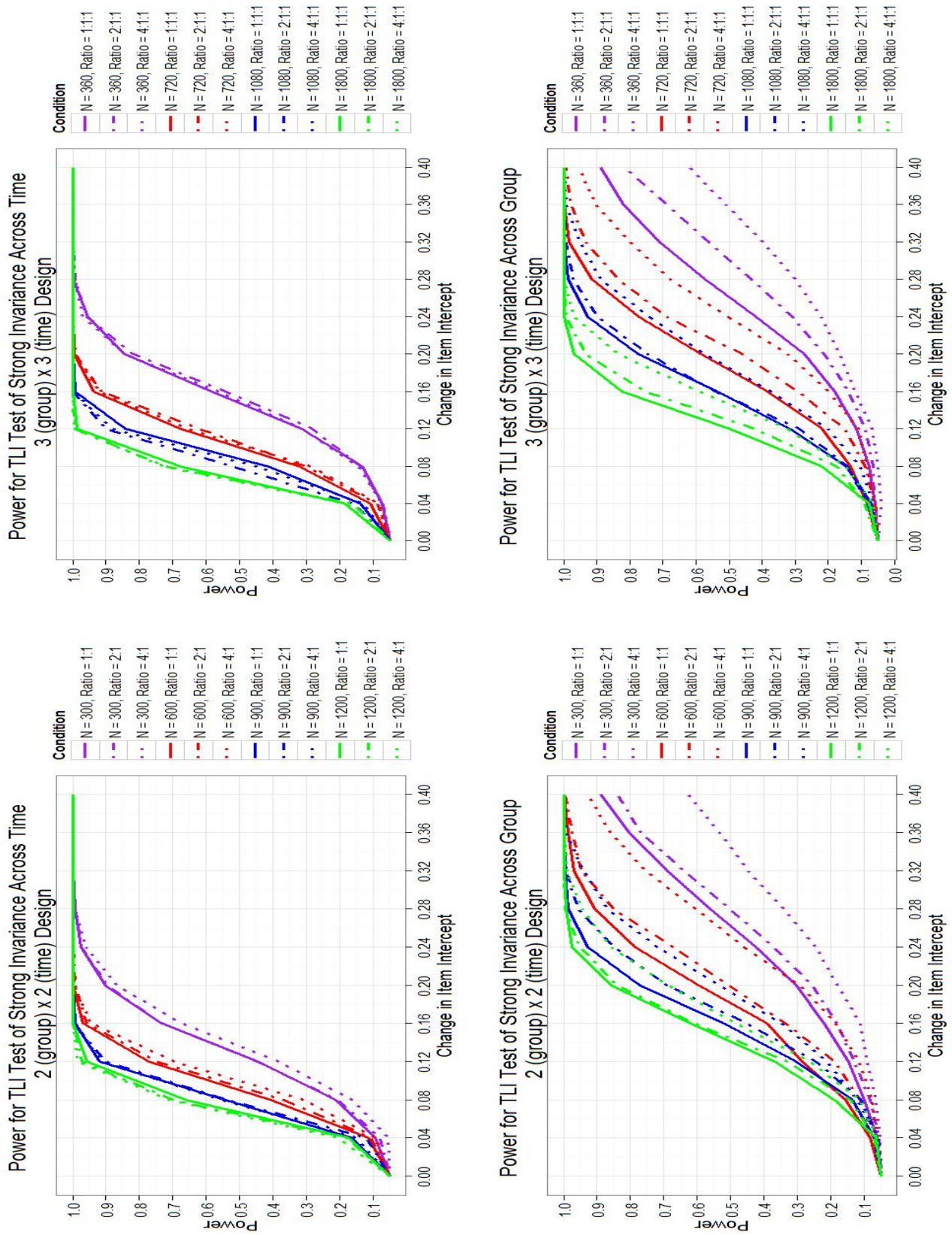


Figure C.7: Power for ΔTLI_A Tests of Weak Invariance

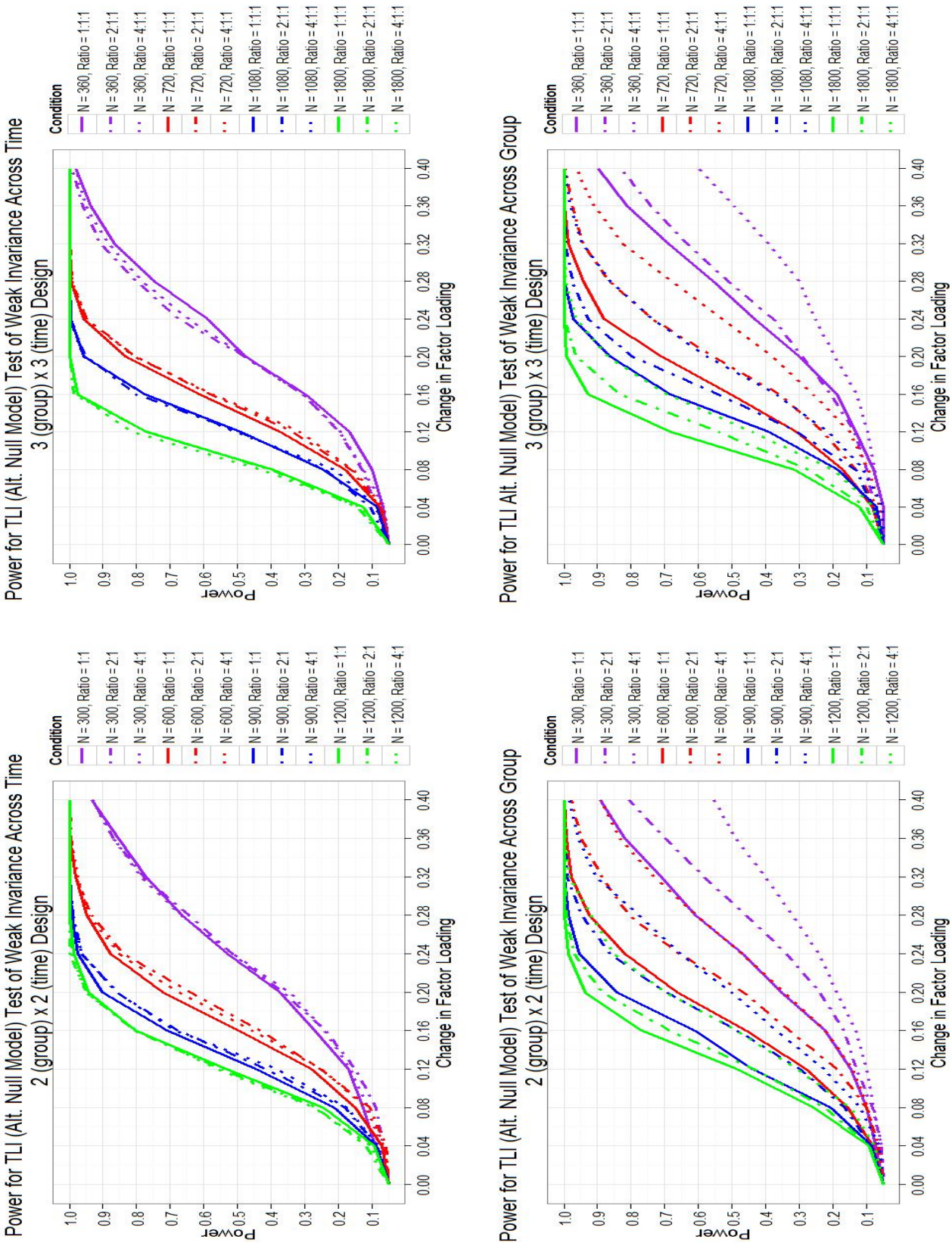
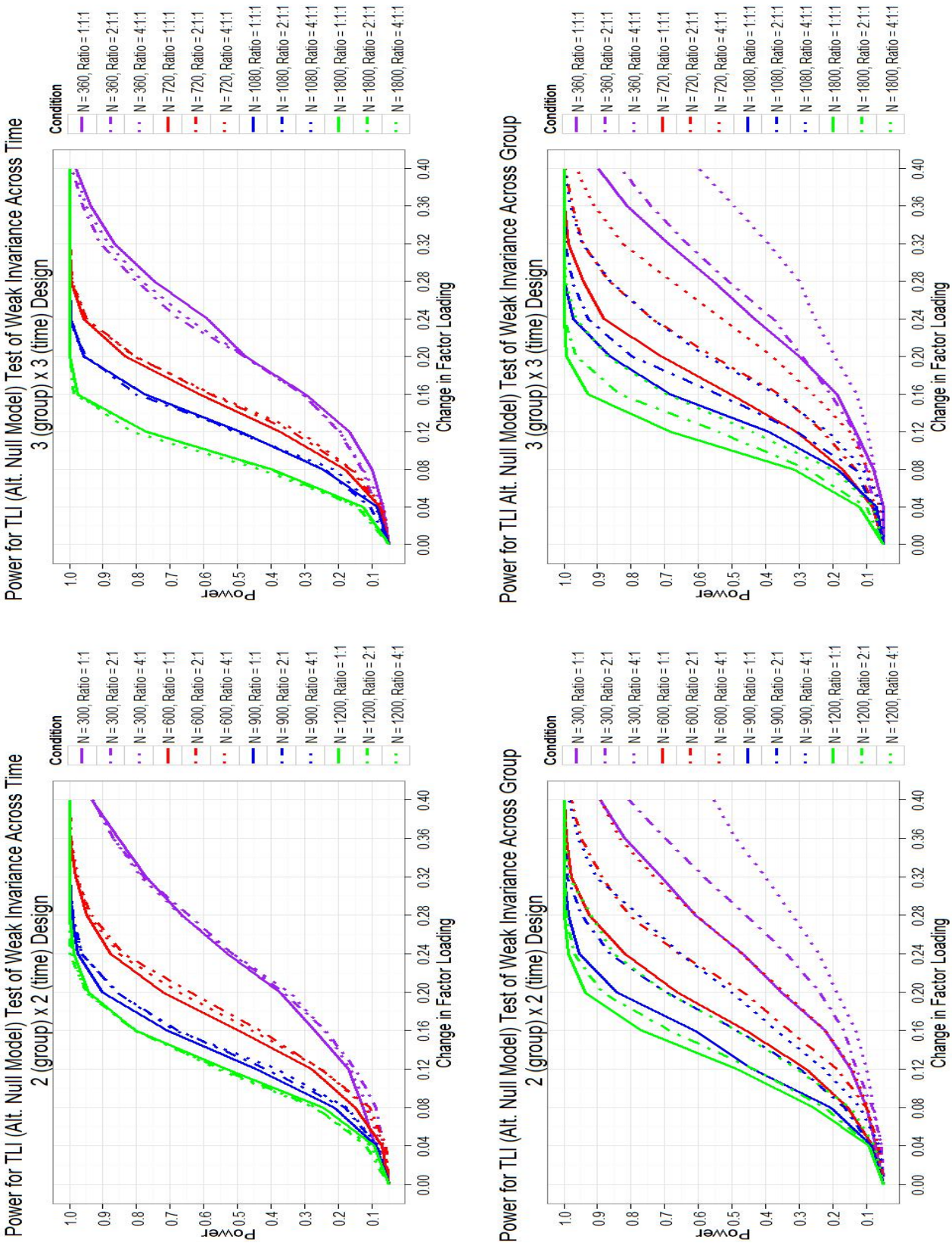


Figure C.8: Power for ΔTLI_A Tests of Strong Invariance



C.5 Results for *AIC*

Figure C.9 displays power of the *AIC* test of weak invariance across group and time for both design types. Figure C.10 displays power of the *AIC* test of strong invariance across group and time for both design types. Overall, larger sample sizes led to greater power to detect changes in both factor loadings and items intercepts. Likewise, as sample size ratio went from unbalanced (e.g., 4:1) to balanced (e.g, 1:1) the the power to detect a change in a factor loading or item intercept increased. The effect of sample size ratio on power was more noticeable in tests of invariance between groups than across time. Interestingly, within both tests of weak and strong invariance across both group and time the 2 (group) x 2 (time) design demonstrated slightly higher power to detect a given effect size than the 3 (group) x 3 (time) design.

For example, within the condition where the $n = 600$ for each group (i.e., $N = 1200$, Ratio = 1:1, and $N = 1800$, Ratio = 1:1:1). Power to detect a $\Delta\lambda = 0.12$ for tests of weak invariance across time was .596 for the 2 (group) x 2 (time) design and .544 for the 3 (group) x 3 (time) design, whereas power for tests of weak invariance across group was .538 for the 2 (group) x 2 (time) design and .414 for the 3 (group) x 3 (time) design. This observed difference was not as pronounced for tests of strong invariance across time, where in the same sample size and ratio condition, power to detect a $\Delta\tau = .08$ was .746 for the 2 (group) x 2 (time) design and .452 for the 3 (group) x 3 (time) design, whereas power for tests of strong invariance across group was .222 for the 2 (group) x 2 (time) design and .078 for the 3 (group) x 3 (time) design.

Figure C.9: Power for *AIC* Tests of Weak Invariance

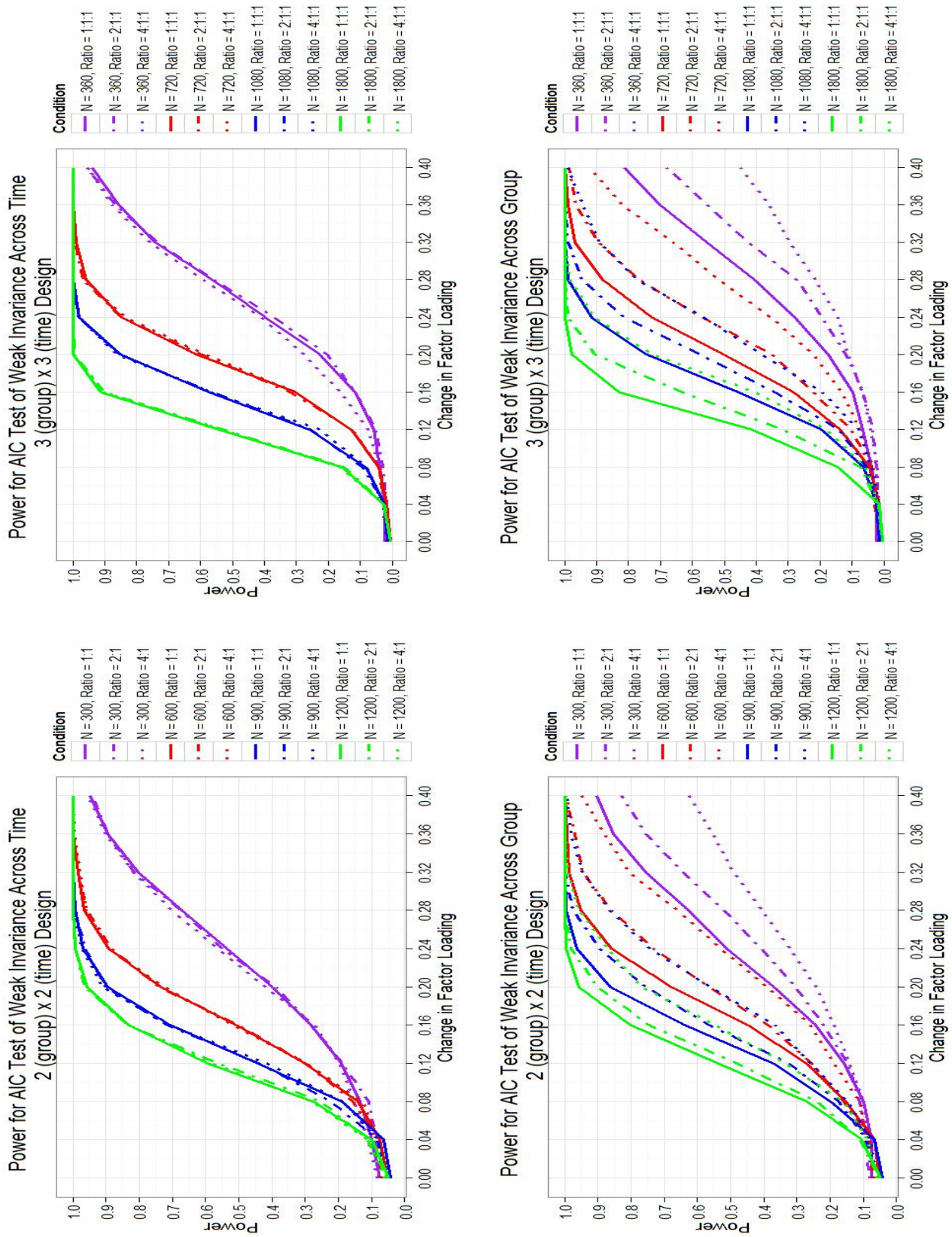
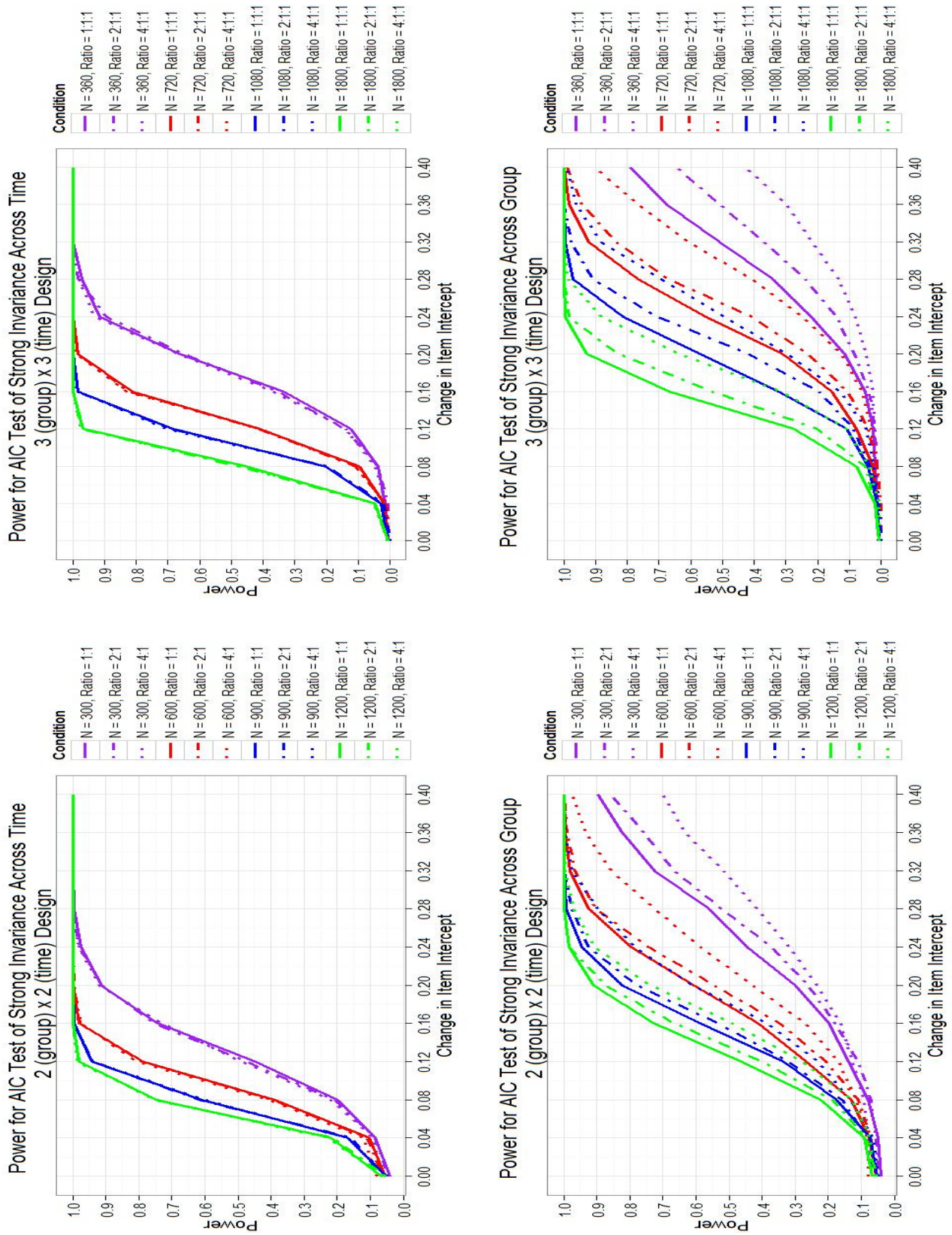


Figure C.10: Power for *AIC* Tests of Strong Invariance



Appendix D

Power using Previously Recommended Cut-offs

D.1 Cheung and Rensvold (2002) Recommendations

Cheung and Rensvold (2002) recommended a cut-off value of $\Delta CFI < .01$ for tests of weak and strong invariance. Figures D.1 and D.2 display power for tests of weak and strong invariance across the current study's conditions using this recommended cut-off.

D.2 Chen (2007) Recommendations

Chen (2007) recommended a cut-off value of $\Delta RMSEA \leq .01$ or $\Delta RMSEA \leq .015$ for tests of weak and strong invariance. Figures D.3 and D.4 display power for tests of weak and strong invariance across the current study's conditions using the larger recommended cut-off value for each case. In addition, Chen (2007) recommended a cut-off value of $\Delta SRMR \leq .025$ or $\Delta SRMR \leq .030$ for tests of weak invariance, and $\Delta SRMR \leq .005$ or $\Delta SRMR \leq .010$ for tests of strong invariance. Figures D.5 and D.6 display power for tests of weak and strong invariance across the current study's conditions using the larger recommended cut-off value for each case.

Figure D.1: Power for $\Delta CFI_A < .01$ Cut-off for Tests of Weak Invariance

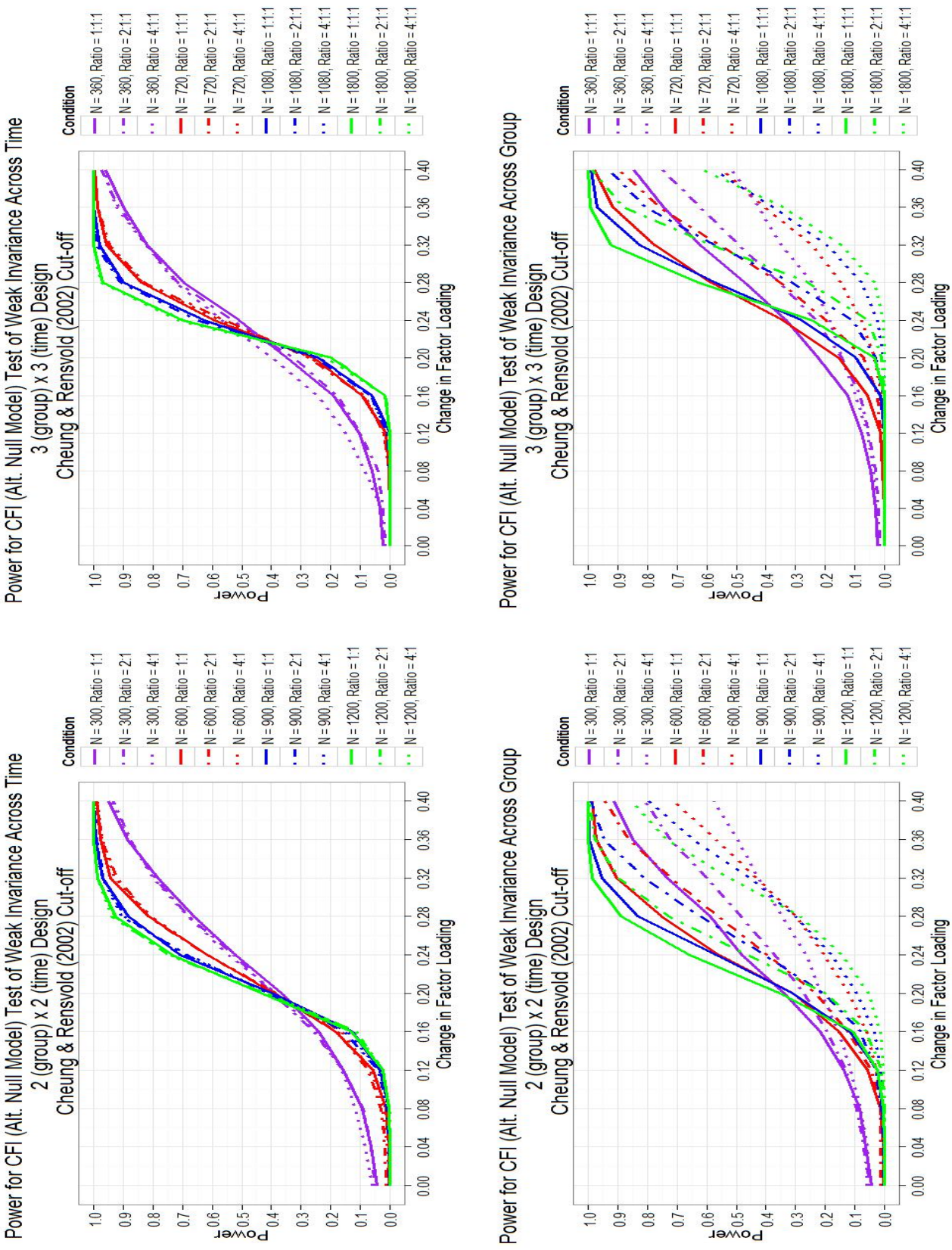


Figure D.2: Power for $\Delta CFI_A < .01$ Cut-off for Tests of Strong Invariance

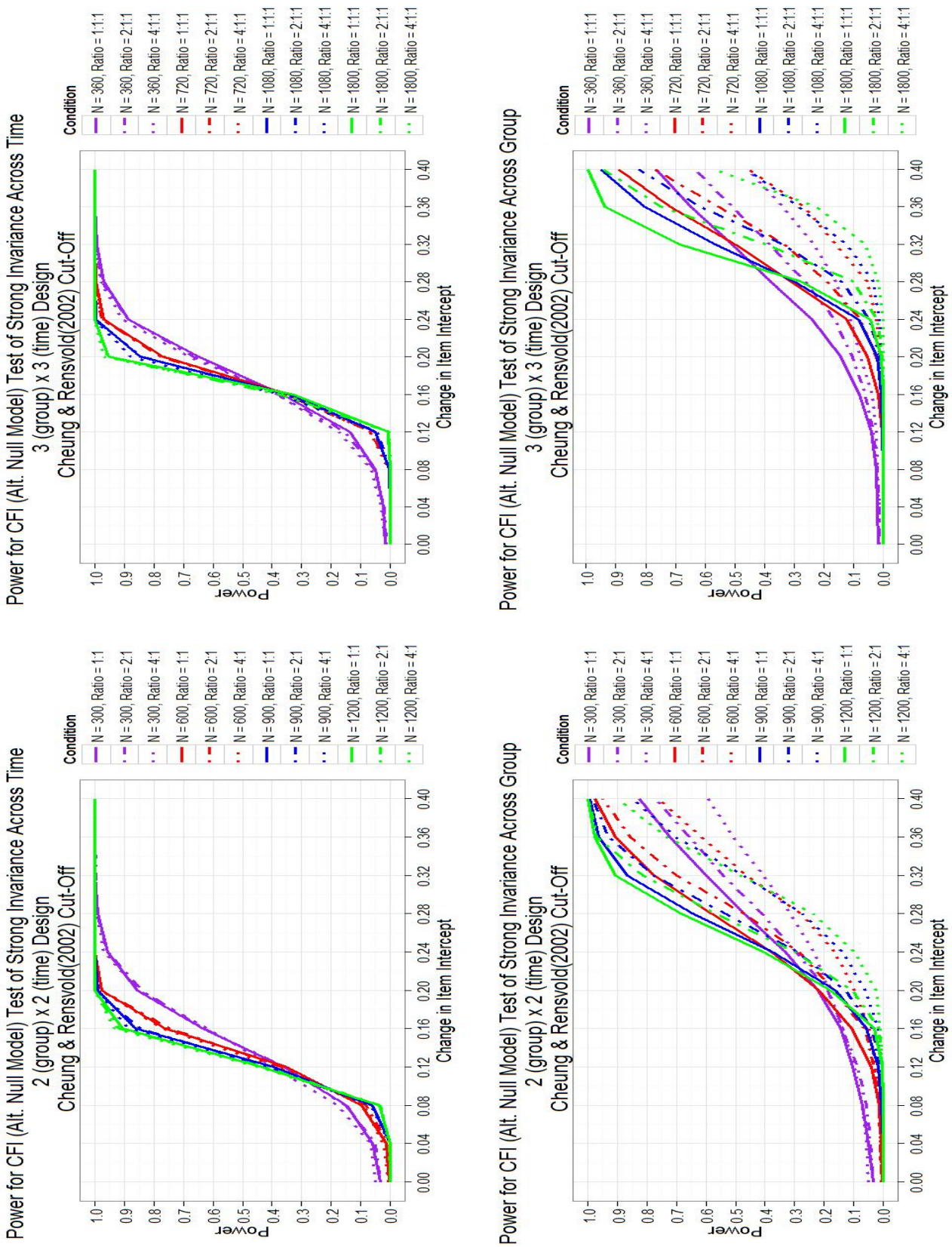


Figure D.3: Power for $\Delta RMSEA \leq .015$ Cut-off for Tests of Weak Invariance

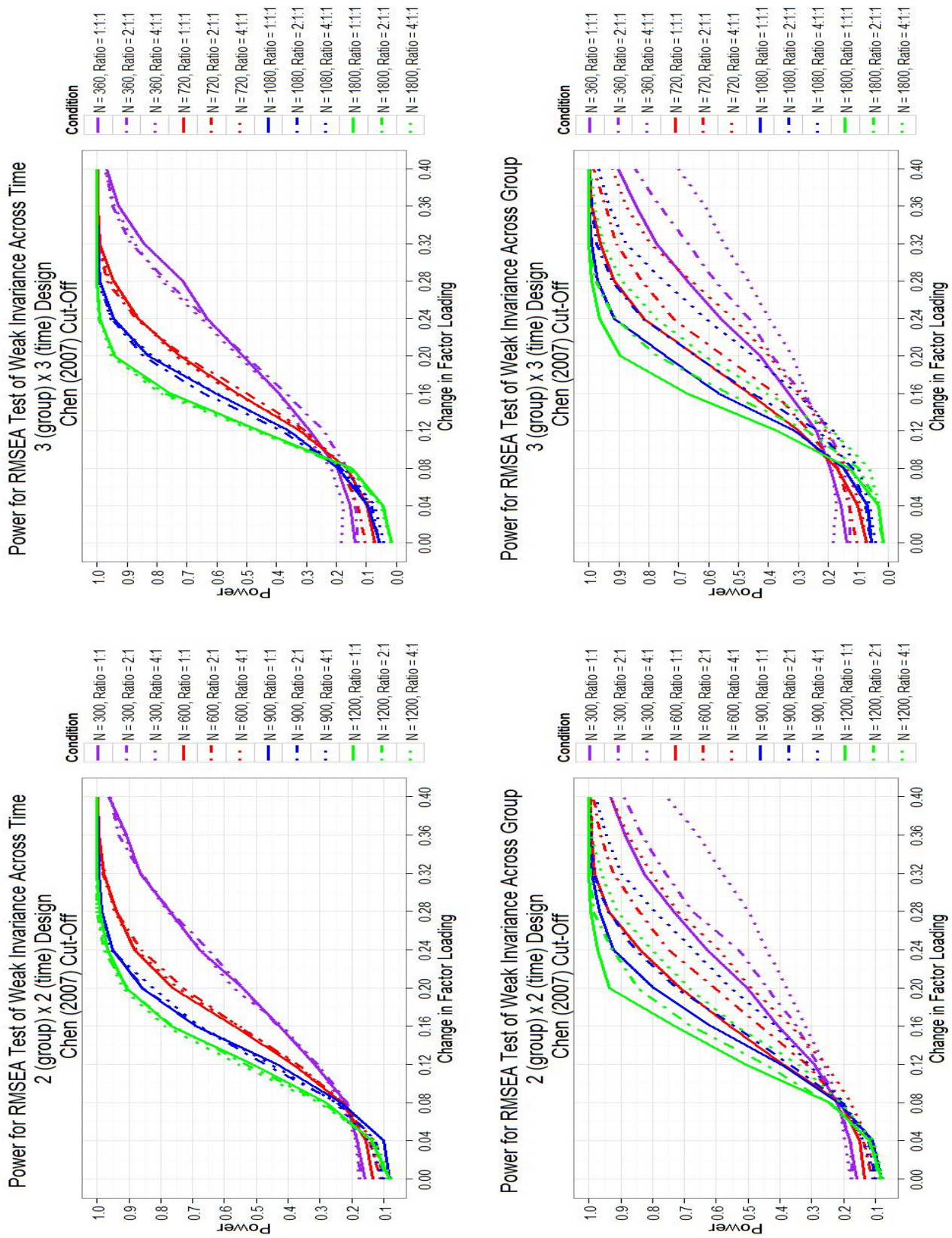


Figure D.4: Power for $\Delta RSMEA < .015$ Cut-off for Tests of Strong Invariance

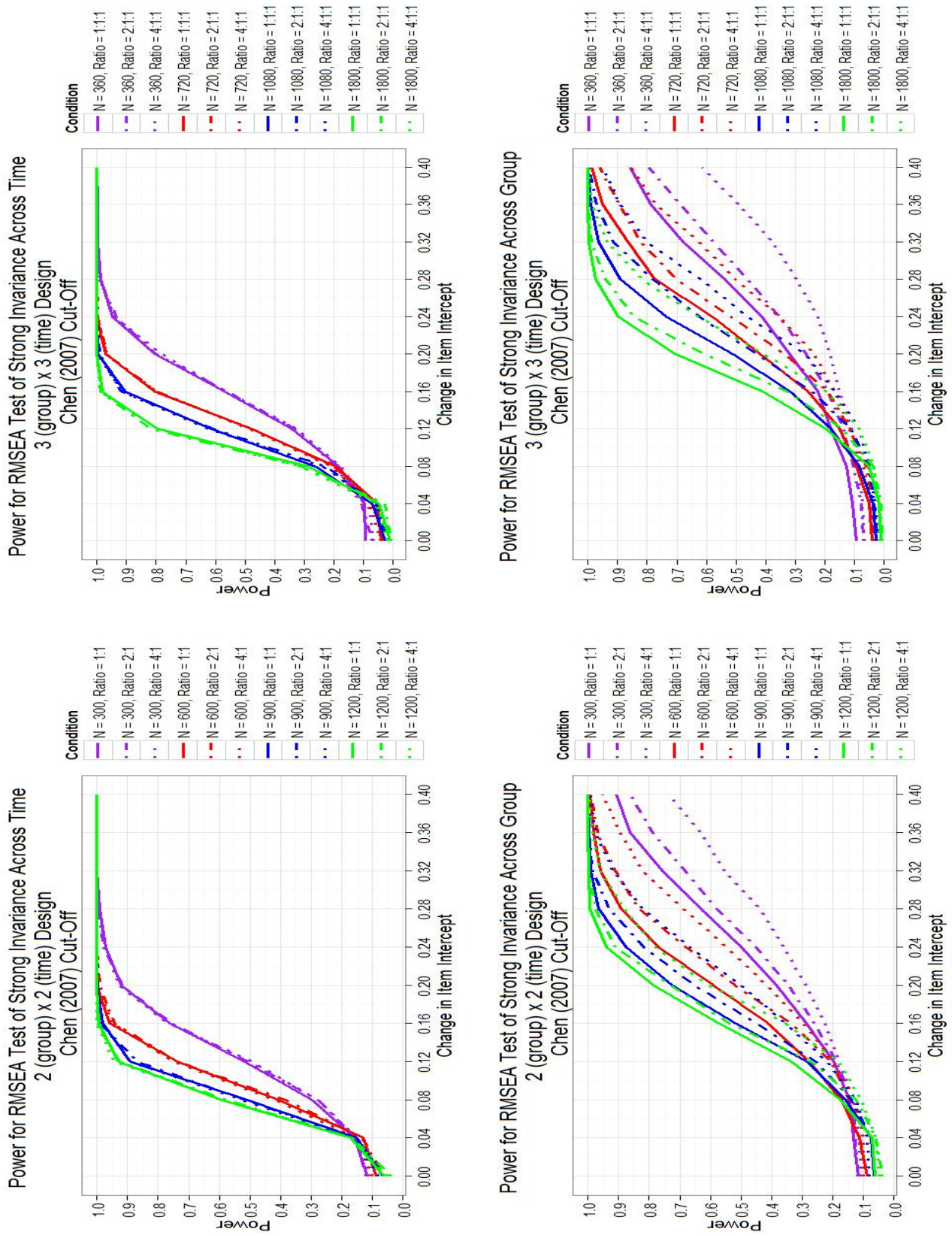


Figure D.5: Power for $\Delta SRMR \leq .030$ Cut-off for Tests of Weak Invariance

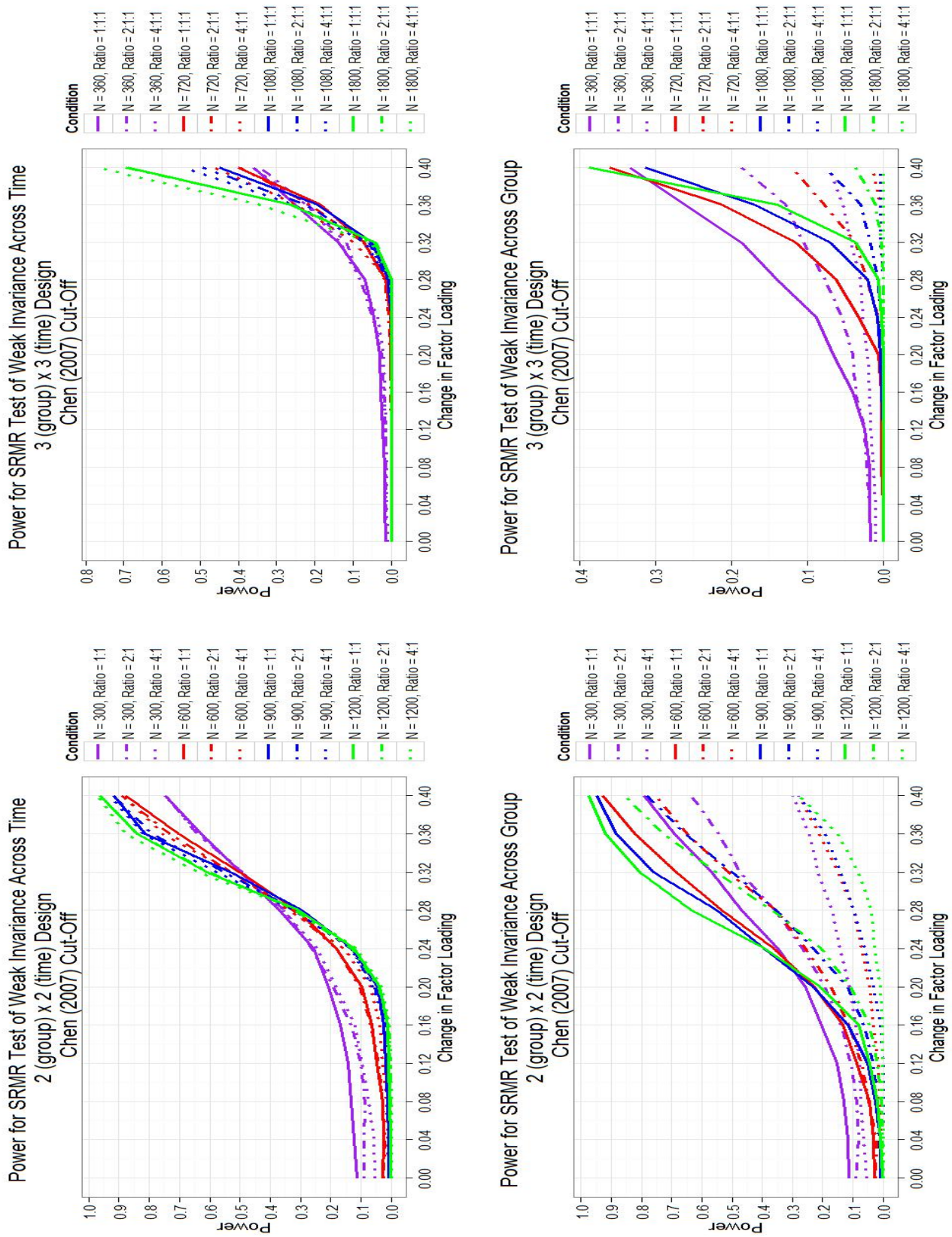
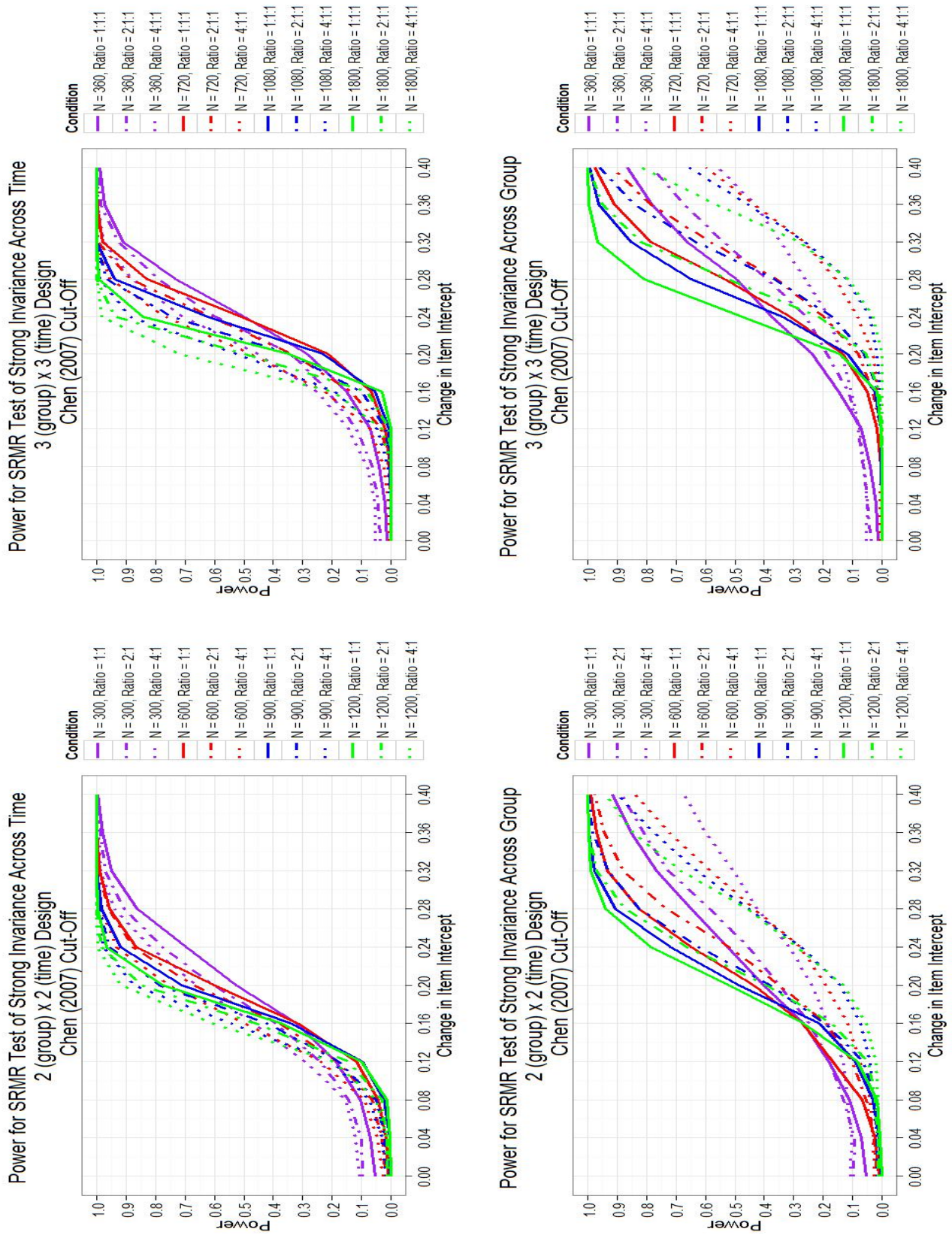


Figure D.6: Power for $\Delta SRMR < .010$ Cut-off for Tests of Strong Invariance



D.3 Meade et al. (2008) Recommendations

Mead et al. (2008) recommended a cut-off value of $\Delta CFI < .005$ for tests of weak invariance and$\Delta CFI < .002$ for tests of strong invariance, nothing their cut-offs were smaller than Cheung and Rensvold's (2002) recommendations because Meade et al. had included many more and more diverse study conditions. Figures D.7 and D.8 display power for tests of weak and strong invariance across the current study's conditions using Meade et al.'s (2008) recommended cut-offs for ΔCFI_A.

Figure D.7: Power for $\Delta CFI_A < .005$ Cut-off for Tests of Weak Invariance

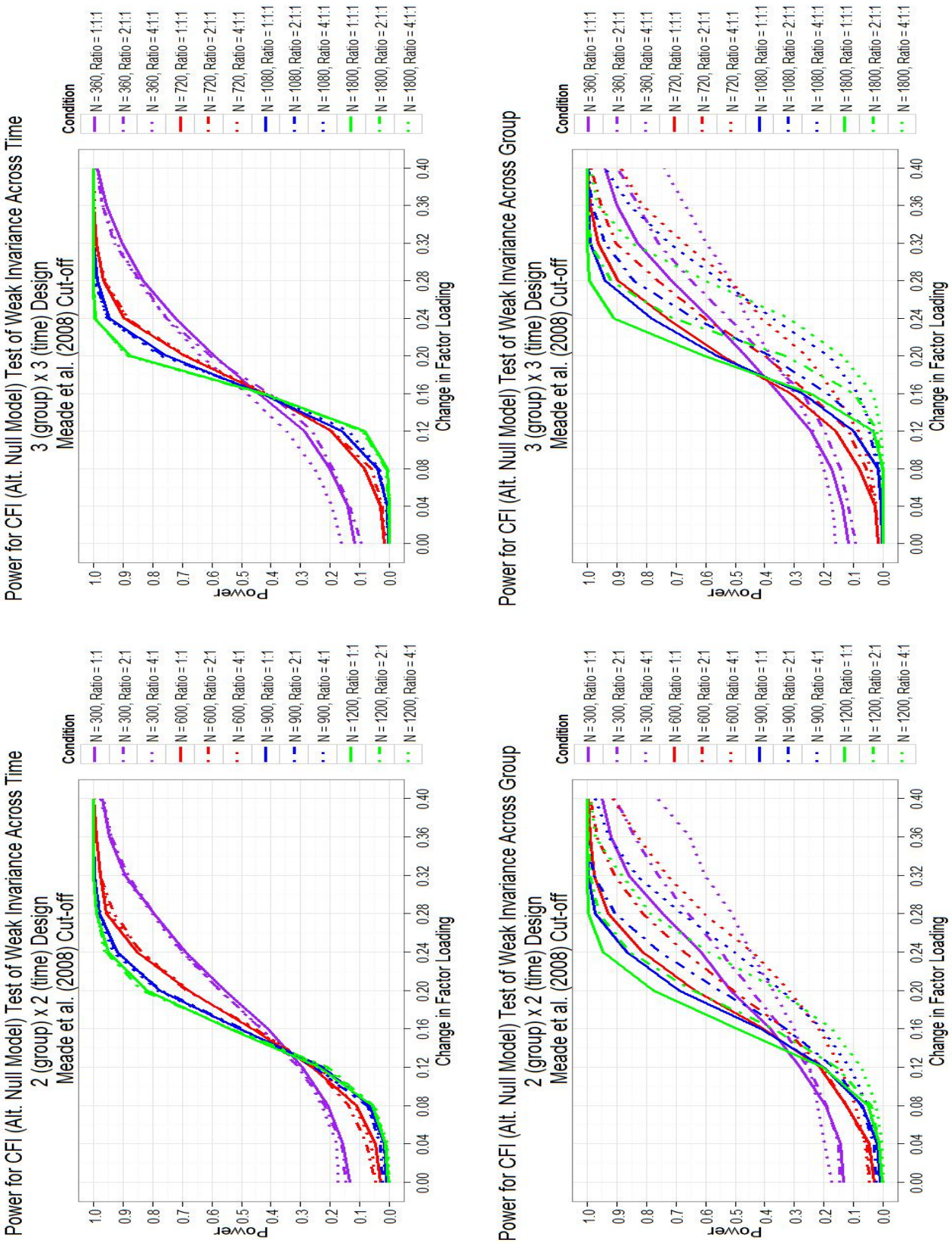


Figure D.8: Power for $\Delta CFI_A < .002$ Cut-off for Tests of Strong Invariance